Essays on invariant item ordering

Ligtvoet, R.

Publication date:
2010

Document Version
Publisher's PDF, also known as Version of record

Essays on Invariant Item Ordering

by Rudy Ligtvoet

ISBN/EAN: 978-90-9025272-8

Essays on Invariant Item Ordering

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit van Tilburg, op gezag van de rector magnificus, prof. dr. Ph. Eijlander, in het openbaar te verdedigen ten overstaan van een door het college voor promoties aangewezen commissie in de aula van de Universiteit op

vrijdag 16 april 2010 om 10.15 uur

door

Rudy Ligtvoet,

geboren op 12 januari 1981 te Hilvarenbeek

Promotores: Prof. dr. K. Sijtsma
Prof. dr. J. K. Vermunt
Copromotor: Dr. L. A. Van der Ark

Acknowledgements

I would like to thank my dissertation supervisors Klaas Sijtsma, Jeroen Vermunt and Andries van der Ark. I learned a lot from all three of you. Klaas, thank you for letting me explore and develop ideas, some of which formed the basis of the chapters in this thesis. Thank you also for your guidance in molding these ideas into papers that people could actually read and understand, and moreover, thank you for your patience. Jeroen, thank you for your advice and enthusiasm. Andries, thank you for your encouragements, support, and looking out for me every so often. Also a special thanks for your naval hospitality.

I am grateful to Janneke te Marvelde and Wicher Bergsma for their invaluable contributions. Larry Hughes and Bas Hemker were kind to provide data. I am also indebted to the committee members for their remarks and suggestions, and I thank the members of IOPS for creating a stimulating environment at the meetings. The Oldendorff Institute supported the research, and I especially thank Ton Heinen and Shirley Baert.

Next, I would like to thank my colleagues at the department of Methodology and Statistics at Tilburg University. Ik wil graag de altijd vriendelijke Marieke Timmermans en Luc Baest bedanken voor de fijne gesprekken. Mijn kamergenoten Hendrik Straat en Judith Conijn, erg bedankt voor jullie tolerantie. Ook dank ik Jacques Hagenaars, Marcel van Assen, Ad Vossen, Marcel Croon, John Gelissen, Guy Moors, Maike Morren, Ruud van Keulen, Johan Braeken, Gabriela Gonzalez Marin, Natalia Kieruj, Peter Kruyen, Daniël van der Palm, and of course I thank Carmen Petrovici, Milos Kankaras, Irena Mikolajun, Fetene Tekle, Roberta Varriale, and Olga Lukociene. Natuurlijk zal ik ook Joost van Ginkel, Marieke Spreeuwenberg en Wouter van der Horst niet vergeten. Ik wil in het bijzonder ook Wilco Emons en Wobbe Zijlstra bedanken voor de vriendschap.

Support also came from old and new friends and I thank you all for your kindness. Thank you Anna Manzoni and Annekathrin Ellersiek, teşekkürler Rahmi İlkılıç, dank jullie wel Gabby Vande Voort, JP Back en Jacobien Kieffer.

Tabii, çok teşekkürler her şey için Süleyman, Gülhas, İlhan, ve Fatma Doğan. Dank ook aan mijn familie, met name Bart van den Broek, Marian en Kees van Dijk, en Linda van Dijk-de Kruijf. Ook verdienen mijn broer Leslie en zus Cindy een speciaal bedankje, al is het maar omdat jullie de beste

broer en zus zijn die ik hebben kan. Daarvoor bedank ik ook mijn vader Ger en moeder Sjaan. Pap en mam, dankt voor jullie steun.

Gönül Doğan Ligtvoet, princess, hayatımın çiçeği, thank you for being my wife, and all that you are.

الحمد لله

Rudy Ligtvoet
Amsterdam, March 4, 2010

Contents

Chapter 1

Introduction

Items in a test or a questionnaire are often scored, such that a higher score is assigned to a more positive response to a statement (personality measurement) or a better solution of a cognitive problem (ability measurement). Items serve as indicators of the attribute of interest. For example, the response to the statement "*I cry at weddings*" may be taken as an indication of the tendency to cry (Vingerhoets & Cornelius, 1990), and higher scores are assigned as the endorsement of the statement is more extreme, for example, on an ordinal rating scale running from "never" to "always". Another example is that scores can be assigned to the subtasks the subject passes in solving a cognitive problem such as arithmetic ability. For example, the item "$\frac{1}{3} + \frac{1}{6} =?$" may be assigned one credit point if the subject finds the common denominator, two credit points if (s)he also correctly adds the numerators, and three credit points if (s)he finally manages to simplify the solution correctly.

The attributes of interest in the examples are not directly observable, but have to be assessed by means of the responses to a set of items that each pertain to a certain aspect of the attribute. Different items can differ in their attractiveness or difficulty. In general we may expect subjects to score lower on the item that states "*I cry when I feel relief*" than on the item "*I cry at weddings*". Likewise, the arithmetic problem "$\frac{2}{3} \div \frac{2}{9} =?$" may generally be considered more difficult than "$\frac{1}{3} + \frac{1}{6} =?$", and thus lead to lower scores. In these examples, the statements with respect to item orderings are expected to hold at the group level meaning that the mean score on item i is smaller than the mean score on item j. However, practical test users often assume that these item orderings also hold for subjects but they do not support this assumption by means of research results. It seems that researchers underestimate the strength of such an assumption or worse, that they do not realize that an ordering relationship that holds at the higher aggregation level of mean item scores does not automatically generalize to the lower lever of individual subjects. Hence, one finds that this aggregation error is regularly made.

Thus, that one item is generally less attractive or more difficult than the

other item does not necessarily mean that this ordering of item difficulties is the same for all subjects. For example, one item may be more difficult than the other for low-ability subjects, while the reverse ordering of item difficulties holds for high-ability subjects. If, however, the ordering is the same for all subjects, then we say that the items are invariantly ordered. Such an *invariant item ordering* (IIO) is convenient for the interpretation and the comparability of test results of different subjects, and is highly relevant in many practical test and questionnaire applications. An example, also to be discussed elsewhere in this thesis, is the use of starting and stopping rules in intelligence testing (e.g., Wechsler, 1999). In child intelligence testing, items are usually administered in the order of increasing difficulty, and the often implicit but not empirically verified assumption is that this difficulty ordering is the same for all children. Based on this assumption, older age groups may skip the easiest items because they are thought to be trivial to them, and children stop solving items when they fail, say, three items in a row. Obviously, the next items are assumed to be too difficult and children should not be bothered with them. Inferences about the subjects' abilities on the basis of this and other adaptive testing strategies rely on the property of IIO, and a violation of the assumption of common item ordering may jeopardize the validity of the inferences made on the basis of such testing procedures. Other examples are the testing of developmental sequences or sequences based on alleged seriousness of symptoms, and person-fit analysis (Sijtsma & Junker, 1996).

To investigate whether IIO holds for a particular test in a particular population of interest, IIO is formally defined within the framework of item response theory (IRT). In IRT models, latent variables represent attributes, and are denoted by θ. The score on an item i is assumed to be a random variable X_i, and concrete item scores or realizations are denoted by $x_i \in \{0, \ldots, m_i\}$. The mean item score for item i is denoted by $E(X_i)$, and interpreted as an index for the difficulty of the item. We index the k items in a test or questionnaire according to increasing item means, so that $i < j$ means that $E(X_i) \leq E(X_j)$. Sijtsma and Hemker (1998) define IIO for the k items in the test as

$$E(X_1|\theta) \leq E(X_2|\theta) \leq \cdots \leq E(X_k|\theta), \qquad \text{(IIO)}$$

for all θ. This definition allow for the possibiliy of ties.

The assumption of IIO obviously is a strong one and difficult to satisfy for a particular test or questionnaire in a particular population. For dichotomously scored items, from the better known IRT models only the Rasch (1960) model and the double monotonicity model (Mokken, 1971) imply IIO but for polytomously scored items none of the well-known IRT models implies the IIO property (Sijtsma & Hemker, 1998). Very little research has been done to develop models implying an IIO and methods that may be used for investigating IIO without the assumption of a particular IRT model (Sijtsma & Hemker, 1998; Sijtsma & Junker, 1996). In our view, this does not mean that the topic

is unimportant but rather that it may be badly understood among practitioners that an item ordering ascertained from the data, either using item means as in classical test theory or latent item location parameters as in item response theory, does not automatically apply to the subject level.

In this thesis, from different angles I tackle the problem of developing models implying IIO and methods that may be used for investigating IIO without the assumption of a particular IRT model. Rather than take one approach and develop it fully, given the no man's land of IIO research I chose to explore different possibilities and make several suggestions for what hopefully will prove to be fruitful approaches.

Arrangement of Chapters

Each of the following chapters starts with an introduction of concepts, definitions, and notation, allowing each chapter to be read independently of the other chapters. In addition, each chapter addresses a particular problem for IIO research, so if one is interested in a specific topic it may pay of to first read the abstracts at the beginning of the chapters.

In Chapter 2, I propose a method for investigating IIO for polytomously scored items, which assesses for each pair of item response functions whether or not they intersect, and I propose a procedure for reducing the large amount of output so as to select a set of items for which IIO holds from a larger set in which IIO does not hold simultaneously for all items. This method is called method *manifest IIO*. Coefficient H^T is defined for polytomously scored items, and given that method manifest IIO supports IIO for a set of items, coefficient H^T expresses the accuracy of this item ordering. A top-down procedure for IIO research by means of method manifest IIO and coefficient H^T is illustrated using a data example.

In Chapter 3, different items are allowed to have different numbers of score categories. For such sets of items, the combination of monotone increasing item response functions (monotonicity, for short) and IIO are investigated simultaneously. A bottom-up procedure is suggested for selecting a set of items for which both monotonicity and IIO hold from a larger set. The procedure is illustrated by an application to data collected by means of an Alzheimers' symptoms checklist.

In Chapter 4, I consider an approach for testing monotonicity and IIO by means of latent class models. The latent variable is approximated by a finite number of discrete latent classes. The Gibbs sampling procedure is used to impose ordinal constraints on the latent class model, where the constraints correspond to the assumptions of monotonicity and IIO. Posterior checks are used to identify the items that do not agree with the constraints corresponding to monotonicity and IIO.

In Chapter 5, the perspective on IIO is such that item response functions are required to be distinguishable in the data before the conclusion of IIO is

drawn. Hence, strict inequality for all θ values is assumed in the definition of IIO. For realistic sample sizes, it is shown that no more than six items can be assumed to fulfill strict IIO. Another result is that for IIO research most subjects need to be sampled form the extremes of the θ distribution, where ironically the fewest observations are located.

In Chapter 6, I define a family of *latent scales* for polytomously scored items. Latent scales are IRT models that imply IIO. The different latent scales are shown to be hierarchically related, and for different levels of the hierarchy testable consequences are derived, allowing the assessment of different definitions of item difficulty ordering. The methodology of Chapter 2 is used to select sets of items that satisfy a particular latent scale from larger sets. Two data examples illustrate the viability of the approach. Finally, in Chapter 7, a summary of the methods and models for IIO is given together with some general conclusions.

Chapter 2

Investigating Invariant Item Ordering for Polytomously Scored Items*

Abstract

This chapter discusses the concept of an invariant item ordering (IIO) polytomously scored items and proposes methods for investigating IIO in real test data. Method *manifest IIO* is proposed for assessing whether or not item response functions intersect. Coefficient H^T is defined for polytomously scored items. Given that IIO holds, coefficient H^T expresses the accuracy of this item ordering. Method manifest IIO and coefficient H^T are used together to analyze a real data set. Topics for future research are discussed.

*This Chapter has been accepted for publication as: Ligtvoet, R., Van der Ark, L. A., Te Marvelde, J. M., & Sijtsma, K. (in press). Investigating an invariant item ordering for polytomously scored items. *Educational and Psychological Measurement.*

2.1 Introduction

In several measurement applications, it is convenient that the items have the same order with respect to difficulty or attractiveness for all subjects. Such an ordering facilitates the interpretation and the comparability of subjects' test results. An item ordering that is the same for all subjects is called invariant item ordering (IIO; Sijtsma & Junker, 1996). Before we define IIO, we first mention several measurement applications in which IIO proves useful.

First, many intelligence tests present the items to children in the order according to ascending difficulty (e.g., Bleichrodt, Drenth, Zaal, & Resing, 1987; Wechsler, 1999). One reason for this presentation order of item administration is to comfort children and prevent them from panicking, which might result from starting with difficult items and which might negatively influence test performance. Another reason is that different age groups are administered different subsets of the items, and subsets of items are more difficult as age increases. For example, the youngest age group starts with the easiest items and a child stops when he or she fails, say, three consecutive items. The next age group always skips the five easiest items, because these items are believed to be trivial to them, and starts at Item 6, and again a child stops when he or she fails, say, three consecutive items. And so on for the next age groups. Several intelligence tests use this administration mode, which assumes that the ordering of the items by difficulty is the same across age groups and children. This assumption usually is ignored in the phase of test construction. In subsequent test use, test practitioners often are unaware that the assumption was never ascertained by means of empirical research, but they use the test as if it were.

Second, several developmental theories assume that abilities or skills go through different phases before they reach maturity (Bouwmeester & Sijtsma, 2007; Raijmakers, Jansen, & Van der Maas, 2004). A simple example is arithmetic ability, for which it may be assumed that development goes through mastering the operation of addition, and then subtraction, multiplication and, finally, division. An arithmetic test, which aims at measuring the degree to which these operations have been mastered, may be assembled and administered such that the hypothesized item ordering by difficulty reflects the assumed ordering of the operations or combinations of the operations. The hypothesized developmental ordering could be investigated using this test with either cross-sectional or, even better, longitudinal data from the population of interest. When the theory proves to be correct, this would lend credence to the diagnostic use of the test and the possibility to pinpoint children's problems with arithmetic as either normal developmental hurdles to be taken or signs of abnormal development.

Third, in attitude and personality testing, and also in the medical context researchers often assume their items to have a cumulative structure, reflecting

a hierarchy of psychological or physical symptoms hypothesized to hold at the subject level (Van Schuur, 2003; Watson, Deary, & Shipley, 2008). For example, in measuring introversion it seems reasonable to expect a higher mean score on a rating scale statement like "*I do not talk a lot in the company of other people*" than on "*I prefer not to see people and do things on my own*", because the latter statement seems to refer to a more intense symptom of introversion. However, an ordering of these statements by group mean scores does not imply that this ordering also holds at the subject level. Indeed, several subjects may indicate a higher prevalence for doing things on their own, but the mixture of the two item orderings may be such that the first still has the highest mean score in the total group. Any set of items can be ordered by means of item mean scores, but whether such an ordering also holds for individual subjects has to be ascertained by means of empirical research. Only when the set of items has IIO, can their cumulative structure be assumed to be valid at the lower aggregation level for subjects.

This study deals with the investigation of IIO for a set of polytomously scored items and extends previous work of Sijtsma and Meijer (1992) and Sijtsma and Junker (1996) for dichotomously scored items. Very little work has been done in this area. Therefore, this study presents some first steps and has an exploratory character. An empirical data example shows that the results may be used for investigating whether IIO holds in sets of polytomously scored items. Finally, directions for future research are discussed.

2.1.1 Definition of Invariant Item Ordering

The context of this study is item response theory (IRT). Let a test contain k polytomously scored items, each of which is characterized by $m + 1$ ordered integer scores. These scores reflect the degree to which a subject solves a complex problem (e.g., a physics problem or a text comprehension problem) or endorsed a statement (e.g., as in Likert items). For $m + 1 = 2$, items are dichotomously scored. Technically, the number of ordered item scores may vary across items but this hampers the comparison of expected item scores for different items. Hence, we follow Sijtsma and Hemker (1998) in only considering equal numbers of ordered item scores; equal numbers are common in many standard tests and questionnaires.

Let random variable X_i denote the score on item i, with realization $x_i \in \{0, \ldots, m\}$. Let θ be the unidimensional latent variable from IRT on which the subjects can be ordered. A test that consists of k items has IIO (Sijtsma & Hemker, 1998) if the items can be ordered and numbered accordingly, such that for expected conditional item scores

$$E(X_1|\theta) \leq E(X_2|\theta) \leq \ldots \leq E(X_k|\theta), \qquad \text{(IIO)}$$

for all θ. IIO allows for the possibility of ties. The expected conditional item

score $E(X_i|\theta)$ is called the item response function (IRF), and IIO implies that the IRFs do not intersect. For dichotomously scored items, $E(X_i|\theta) = P(X_i = 1|\theta)$, which is the conditional probability of a correct score or the endorsement of a statement.

IIO is a strong requirement in measurement practice. Researchers sometimes assume that a fitting IRT model implies that the items have the same ordering by difficulty or attractiveness for all subjects but this assumption requires modification. For dichotomously scored items, Sijtsma and Junker (1996) showed that only IRT models that employ nonintersecting IRFs imply IIO. Examples are the Rasch (1960) model and the Mokken (Mokken & Lewis, 1982) double monotonicity model, but the much-used two- and three-parameter logistic models (Birnbaum, 1968), which allow intersecting IRFs, do not imply the IIO property. For polytomously scored items, Sijtsma and Hemker (1998) showed that popular IRT models such as the partial credit model (Masters, 1982), the generalized partial credit model (Muraki, 1992), and the graded response model (GRM; Samejima, 1969) do not imply an IIO. Thus, when any of these models give an accurate description of the data, one cannot conclude that the items follow the same ordering by difficulty or popularity for each subject from the population of interest. Sijtsma and Hemker (1998) showed that only restrictive polytomous IRT models, such as the rating scale model (Andrich, 1978), a rating scale version of Muraki's (1990) restricted GRM, and the isotonic ordinal probabilistic model (Scheiblechner, 1995) imply IIO.

Thus, there appears to be a mismatch between many of the IRT models for polytomously scored items and the IIO property. This mismatch is due to an aggregation phenomenon, which is illustrated by means of the GRM and a special case of this model. We assume a unidimensional latent variable θ, and item scores that are locally independence. Response functions of polytomously scored items are defined for separate item scores and given that an item has $m + 1$ different scores, for each item m such response functions are needed (Mellenbergh, 1995; Molenaar, 1983). An example of these response functions are the item step response functions (ISRFs) of the class of cumulative probability models, which are defined by the conditional probability $P(X_i \geq x_i|\theta)$, for $x_i = 0, \ldots, m$; by definition, $P(X_i \geq 0|\theta) = 1$ and $P(X_i \geq m + 1|\theta) = 0$.

Given IIO, one is interested in statistical information at the aggregation level of the item rather than the level of the item scores. Hence, we considering the IRF, which is related to the m ISRFs by means of

$$E(X_i|\theta) = \sum_{x_i=1}^{m} P(X_i \geq x_i|\theta). \tag{1}$$

Sijtsma and Hemker (1998) used relationships like this one to show that for many polytomous IRT models, combining the m ISRFs of items into IRFs, does not result in IIO. These authors also showed that one needs restrictions on the

relationships between the ISRFs of different items in the test or questionnaire to obtain IIO. Two examples of relationships between ISRFs and IRFs are given, one resulting in failure of IIO and the other in IIO.

First, in Samejima's (1969) GRM each item has m threshold parameters such that $\beta_{1_i} \leq \ldots \leq \beta_{m_i}$ (i.e., the m ISRFs have a fixed order). In addition, each item has one discrimination parameter $\alpha_i > 0$; then, the ISRF for item score x_i is defined as

$$P(X_i \geq x_i|\theta) = \frac{e^{\alpha_i(\theta-\beta_{x_i})}}{1+e^{\alpha_i(\theta-\beta_{x_i})}}, \tag{2}$$

for item scores $x_i = 1, \ldots, m$. The summation across the m ISRFs in Equation 2 across the m item scores yields IRF $E(X_i|\theta)$ (Equation 1). As an example for the failure of IIO for the GRM, consider two items: for item 1 we have $\beta_{x_1} = (-1.5, 1.5)$ and $\alpha_1 = 1$, and for item 2 we have $\beta_{x_2} = (.5, .5)$ and $\alpha_2 = 3$. Figure 2.1a shows the ISRFs of these two items. Using Equations 1 and 2, it is shown in Figure 2.1c that the two IRF intersect at $\theta = 0$; hence IIO does not hold.

Second, restricting the slope parameters of Muraki's (1990) rating scale version of the GRM (Sijtsma & Hemker, 1998) places restrictions on the relationships of the ISRFs of different item, which results in IIO. Let α denote a general discrimination parameter, λ_i an item specific location parameter, and ϵ_x is the distance of the x_ith ISRF to location λ_i, so that $\beta_{x_i} = \lambda_i + \epsilon_x$, and with the restriction that $\sum \epsilon_x = 0$; then, the x_ith ISRF of item i is

$$P(X_i \geq x_i|\theta) = \frac{e^{\alpha(\theta-\lambda_i-\epsilon_x)}}{1+e^{\alpha(\theta-\lambda_i-\epsilon_x)}}, \tag{3}$$

for item scores $x_i = 1, \ldots, m$. All items show the same dispersion of the ISRFs around the location parameters λ_i. For two items satisfying Equation 3, Figure 2.1b and 2.1d show that they have IIO.

Two sources of confusion seem to exist with respect to IIO. The first is that if an IRT model does not imply IIO, then the IIO property cannot be important. We emphasize that it is the measurement application which determines whether IIO is important, not the IRT model. If a particular IRT model does not give information about IIO, other methods have to be used in data analysis for ascertaining whether IIO is valid. The second source of confusion is that the IIO property applies to particular content areas but not to others, and that it applies to rating scale items but not to constructed-response items. The examples given in the beginning of the chapter illustrated that IIO may be important in different content areas. This is also true for different item types. For example, in intelligence tests many items require constructed responses, as in explaining to the test administrator the use of a particular object (e.g., a hammer, a car). If such items are administered in an ascending difficulty

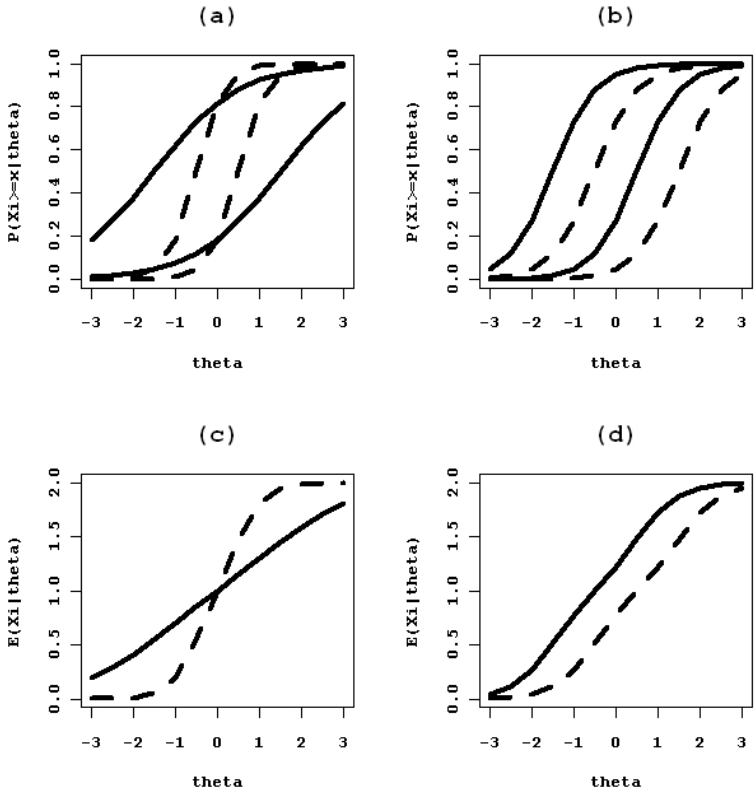

Figure 2.1: (a) Two items having two ISRFs under the GRM, (c) failing IIO, and two items having two ISRFs under the restricted rating scale version of the GRM, (d) having IIO.

ordering, IIO is assumed, which has to be supported by empirical research.

2.1.2 Investigating Invariant Item Ordering

In IIO research for polytomously scored items, a distinction is made between IRFs that are close together and IRFs that are further apart. If IRFs are close together, subjects produce data that contain little information about the item ordering, resulting in an inaccurate ordering. If IRFs are far apart, subjects produce data that contain more information on the actual item ordering. Thus, given that IIO has been established, an index for the distance between the IRFs can be interpreted as an index of the accuracy of the ordering of the IRFs. In this study, the IRFs of the k polytomously scored items, defined by $E(X_i|\theta)$, are estimated, and IIO is ascertained, and if it is concluded that IIO holds, a generalization of coefficient H^T is used, proposed by Sijtsma and Meijer (1992)

for dichotomously scored items, to express the accuracy of the item ordering.

Sijtsma and Meijer (1992) demonstrated by means of a simulation study that for k invariantly ordered dichotomously scored items coefficient H^T increased as the mean distance between the item locations increased, or as the item discrimination increased (both manipulations have the effect that IRFs are further apart), while other properties of the IRFs and the distribution of θ were kept constant. They did not find convincing support for different values of H^T to distinguish between the situations for which IIO held or did not hold (yet suggested tentative rules of thumb for making this distinction; to be discussed later). In a pilot study we found that this distinction was even more difficult to make for polytomously scored items.

In what follows, a two-step procedure for IIO research is proposed for polytomously scored items. First, we discuss the estimation of the IRFs, and propose method *manifest IIO*, which is based on the estimation of IRFs for dichotomously scored items (see Molenaar & Sijtsma, 2000, pp. 74-78) and which evaluates for each pair of IRF estimates whether or not they intersect. By means of a simulation study, the Type I error rate and power of the method was investigated. Second, coefficient H^T for polytomously scored items is discussed. By means of a computational study, the influence of different item and test properties on H^T was investigated under the situation for which IIO holds. Finally, method manifest IIO and coefficient H^T were used to analyze a real data set.

2.2 Method Manifest IIO

2.2.1 Estimation of IRFs and Pairwise Inspection of IIO

Method manifest IIO is available from the R package `mokken` (Van der Ark, 2007). Let $Y_{ij} = X_+ - X_i - X_j$ be the rest score; that is, the sum score on all items excluding the items i and j. Also, y be the realization of Y_{ij}, with $y = 0, \ldots, m(k-2)$. Let $E(X_i|Y_{ij})$ be the estimated IRF of item i. If population item means are ordered such that for item pair i and j, $E(X_i) \leq E(X_j)$, then IIO implies that

$$E(X_i|\theta) \leq E(X_j|\theta), \tag{4}$$

for all θ. Ligtvoet, Van der Ark, Bergsma, and Sijtsma (2009) showed that Equation 4 implies that

$$E(X_i|y) \leq E(X_j|y), \tag{5}$$

for all $Y_{ij} = y$.

Equation 5 is investigated for each pair of items using conditional sample means $\overline{X}_{i|y}$ and $\overline{X}_{j|y}$, for all y. If it is found that $\overline{X}_{i|y} > \overline{X}_{j|y}$ (i.e., a violation of IIO occurs), a one-sided t test is used to test the null hypothesis that $E(X_i|y) = E(X_j|y)$ against the alternative that $E(X_i|y) > E(X_j|y)$. Rejection of the null

hypothesis for at least one value of y leads to the conclusion that items i and j are not invariantly ordered. If the number of subjects having a rest score y is too small for accurate estimation, adjacent rest score are combined into rest-score groups until the group size exceeds a preset minimum (Molenaar & Sijtsma, 2000, p. 67). A protection against taking very small violations seriously is to test sample reversals only when they exceed a minimum value denoted *minvi*. Molenaar and Sijtsma (2000, pp. 67-70) recommend for dichotomously scored items the default value $minvi = .03$. Polytomously scored items have a greater score range, and a logical choice for *minvi* is $m \times .03$. Whether this is a reasonable choice was investigated in a simulation study (next section).

The following sequential procedure is used for method manifest IIO. First, for each of the k items the frequency is determined the item is involved in a significant violation that exceed *minvi*. If none of the items is involved in violations, we conclude that IIO holds for all k items; else, the item with the highest frequency is removed from the test. Second, the procedure is repeated for the remaining $\binom{k-1}{2}$ item pairs, and if an item is removed, for the remaining $\binom{k-2}{2}$ item pairs, and so on. When several items have the same frequency of significant violations, the items having the smallest scalability coefficients (Sijtsma & Molenaar, 2002, p. 57) may be removed but researchers may also consider other exclusion criteria, such as item content.

This procedure is suited for exploratory data analysis but for confirmatory purposes, when one wants to know whether all k items have an IIO, manifest IIO is checked for all item pairs but items are not removed. For the set of items for which no significant violation is found coefficient H^T is computed to evaluate the degree of accuracy of the item ordering. Coefficient H^T is discussed in the next section.

2.2.2 Type I Error and Power of Method Manifest IIO

A Monte Carlo study was performed to investigate the Type I error rate (probability that IIO is incorrectly rejected) and the power (probability that IIO is correctly rejected) of method manifest IIO. Here, we only considered the confirmatory procedure for assessing IIO for all k items.

Method

The following six design factors were used: 1) Situation IIO or not IIO, 2) Size of *minvi*, 3) Item discrimination, 4) Sample Size, 5) Test Length, and 6) number of Score Categories.

1. *Situation IIO or not IIO.* Samejima's (1969) GRM (Equation 2) was used to generate data for those cells of the design for which IIO did not hold. Particular choices of item parameters may coincidentally produce IIO. A pilot study showed, however, that IRFs almost always intersected in

dense regions of the latent variable, so that it seemed safe to use the GRM for the situation in which IIO did not hold. The restricted rating scale version of the GRM (Equation 3) was used to generate data for the cells of the design that correspond to the situation in which IIO did hold. Item parameters β_{x_i} were sampled from $N(0,1)$, where in case of Equation 3 $\lambda_i = \frac{1}{m}\sum_x \beta_{x_i}$, and $\epsilon_x = \frac{1}{k}\sum_i \beta_{x_i}$. Slope parameter are discussed below.

2. *Size of minvi.* A total of 16 values for *minvi* were chosen to covering a wide range: $minvi = .00(.03).45$. These increments include the suggestion that $minvi = m \times .03$. The value $minvi = .00$ implies that all violations, however small, were tested.

3. *Item discrimination.* Two levels of the item discrimination were in Equations 2 and 3, corresponding to weak and moderate discriminating items. For the situation in which IIO did not hold for weak discriminating items, parameters α_i were sampled form $\log N(-.5\ln 20, \ln 5)$, which yield $E(\alpha_i) = .5$ and $Var(\alpha_i) = 1$. For moderate discriminating items, these parameters α_i were sampled form $\log N(-.5\ln 2, \ln 2)$, which yield $E(\alpha_i) = 1$ and $Var(\alpha_i) = 1$. For the IIO situation a single slope parameter was sampled from the same distributions for weak and moderate item discriminations.

4. *Sample Size.* Sample Size had three levels: $n = 200, 433$, and 800, where $n = 433$ corresponds to the data example below. Values of θ were sampled form $N(0,1)$.

5. *Test Length.* Test Length had three levels: $k = 5, 10$, and 15, where $k = 10$ corresponds to the data example.

6. *Score Categories.* Three numbers of Score categories were considered: $m + 1 = 3, 5$, and 7, where $m + 1 = 5$ corresponds to the data example.

For each of the (Situation × Item Discrimination × Sample Size × Test Length × Score Categories =) 108 cells of the design 500 data sets were generated and each data set was analyzed by means of method manifest IIO for each of the 16 *minvi* values. The Type I error rate was computed for each of the cells corresponding to IIO, and the power was computed for each of the cells corresponding the not IIO situation.

Results

The Type I error rate of method manifest IIO ranged from .000 to .725 across all design cells (mean = .151, standard deviation = .195), and the power ranged from .013 to 1.000 (mean = .686, standard deviation = .337). Only significant main effects on the Type I error rate and power are discussed (Kruskal-Wallis test for several independent samples, nominal Type I error of .05).

Table 2.1: Type I Error Rate and Power of Method Manifest IIO for Different Sizes of *minvi*.

minvi	Discrimination			
	Weak		Moderate	
	Type I Error	Power	Type I Error	Power
.00	.350	.924	.098	.998
.03	.350	.924	.098	.998
.06	.350	.924	.098	.998
.09	.350	.924	.098	.998
.12	.350	.924	.098	.998
.15	.350	.924	.098	.998
.18	.348	.924	.098	.998
.21	.346	.918	.092	.998
.24	.284	.892	.076	.998
.27	.170	.850	.040	.996
.30	.100	.776	.026	.986
.33	.058	.718	.006	.980
.36	.030	.654	.004	.962
.39	.012	.596	.000	.950
.42	.002	.558	.000	.924
.45	.000	.490	.000	.902

Table 2.1 shows the Type I error rate and power for the two levels of Slope of the IRFs and the 16 levels of minvi, for $n = 433$, $k = 10$, and $m + 1 = 5$ (choices corresponded to data example). For these cells of the design, Type I errors were higher for weak discriminating items and small *minvi*-values, and decreased as either *minvi* increased or item discrimination increased. Power was generally high, except for weak item discrimination in combination with high *minvi*-values. The reason that the results are the same for all *minvi*-values lower than .18 is that the critical value of the t-test is larger than these values of *minvi*. Based on these results, the suggestion for $minvi = m \times .03$ seems reasonable if items from the data example show at least a moderate discrimination.

Across the cells of the design (averages are reported here) an increase in *minvi* resulted in lower Type I error rates: .240 ($minvi = .00$) and .030 ($minvi = .45$) and lower power: .790 ($minvi = .00$) and .490 ($minvi = .45$). The Type I error rates were higher for weak discriminating items than for moderate discriminating items: .211 and .092, respectively, whereas the power was lower for weak discriminating items than for moderate discriminating items: .614 and .758, respectively. Larger Sample Size resulted in lower Type I error

Table 2.2: Summary of Main Effects on Type I Error Rate and Power of Method Manifest IIO.

	Type I Error	Power
Size of *minvi*	–	–
Slope of the IRFs.	–	+
Sample Size	–	·
Test Length	+	+
Score Categories	+	+

rates: .269 ($n = 200$) and .085 ($n = 800$). Larger Test Length resulted in larger Type I errors: .021 ($k = 5$) and .285 ($k = 15$), and larger power: .350 ($k = 5$) and .913 ($k = 15$). Finally, the number of Score Categories increased the Type I errors: .125 ($m + 1 = 3$) and .162 ($m + 1 = 7$), and increased power: .486 ($m+1 = 3$) and .827 ($m+1 = 7$). Table 2.2 gives an overview of the significant positive (+) and negative (–) main effects.

Discussion

Higher *minvi*-values result in a smaller Type I error rate but also lowered the power of method manifest IIO. The choice of *minvi* thus depends on the specific application for which IIO is investigated. A high cost of incorrectly accepting IIO requires a lower *minvi*-value, but in other cases, including our data example, $minvi = m \times .03$ may be appropriate. Method manifest IIO also benefits from higher discriminating items, and larger sample sizes.

2.3 Polytomous Coefficient H^T

2.3.1 Definition of Coefficient H^T

Let $\mathbf{X}$ denote the data matrix of n subjects (rows) by k items (columns), with scores $x = 0, \ldots, m$ in the cells. Coefficient H (Mokken & Lewis, 1982; Sijtsma & Molenaar, 2002, chap. 4) is a measure for the accuracy by which k items constituting a scale for ordering subjects (Mokken, Lewis, & Sijtsma, 1986). Sijtsma and Meijer (1992) showed for dichotomous items that when H is computed on the transposed data matrix, the coefficient denoted H^T, is a measure for the accuracy by which n respondents order k items. Here, we generalize coefficient H^T to polytomously scored items.

Let subjects be indexed by g and h, and let the vector $\mathbf{X}_g$ ($g \in \{1, \ldots, n\}$) contain the scores of subject g on the k items. We assume that the k item scores show at least some variation, so that $\text{var}(\mathbf{X}_g) > 0$ for all g. Let $\text{cov}(\mathbf{X}_g, \mathbf{X}_h)$ be the covariance between the scores of subjects g and h, and $\text{maxcov}(\mathbf{X}_g, \mathbf{X}_h)$

the maximum possible covariance given the marginal distributions of the item scores the subjects. The sum score on item i is denoted by $Y_i = \sum_g X_g$. Let vector $\mathbf{Y}$ contain the k item sum scores, and let vector $\mathbf{Y}_g = \mathbf{Y} - \mathbf{X}_g$ contains the k sum scores excluding the contribution of subject g. The subject scalability coefficient H_g^T is defined as the weighted normalized covariance,

$$H_g^T = \frac{\sum\limits_{h \neq g} \text{cov}(\mathbf{X}_g, \mathbf{X}_h)}{\sum\limits_{h \neq g} \text{maxcov}(\mathbf{X}_g, \mathbf{X}_h)} = \frac{\text{cov}(\mathbf{X}_g, \mathbf{Y}_g)}{\text{maxcov}(\mathbf{X}_g, \mathbf{Y}_g)}. \tag{6}$$

Thus, coefficient H_g^T expresses the association between the k item scores of subject g and the k sum scores of the remaining subjects. Because even for small samples, $\mathbf{Y} \approx \mathbf{Y}_g$, coefficient H_g^T expresses the degree to which the scores of subject g have the same ordering as the sum scores.

When IIO holds for the k items, theoretically we expect a positive association between the ordering of the item scores in $\mathbf{X}_g$ and the total scores $\mathbf{Y}_g$. When IRFs are close together, we expect this ordering of the item scores to be unstable and the values of many coefficients H_g^T to be low. When IRFs are further apart, we expect the orderings of the item scores to be more stable and better in agreement with the ordering of the item totals, thus resulting in many higher H_g^T values. Coefficient H^T wraps up the n subject coefficients as

$$H^T = \frac{\sum\limits_{g} \text{cov}(\mathbf{X}_g, \mathbf{Y}_g)}{\sum\limits_{g} \text{maxcov}(\mathbf{X}_g, \mathbf{Y}_g)}. \tag{7}$$

When k items have IIO, the value of coefficient H^T is higher the further the IRFs are apart.

For k invariantly ordered items, assuming LI it follows that $0 \leq \min H_g^T \leq H^T \leq \max H_g^T \leq 1$ (see Appendix A). The value of 0 is obtained if the k IRFs coincide and $\text{cov}(\mathbf{X}_g, \mathbf{X}_h) = 0$ for all subject pairs. Maximally, $H^T = 1$, and this value is obtained if the agreement between the subjects' ordering of item scores and the ordering of the corrected item totals is maximal. We used a computational study to investigate the influence of item and test properties on the value of coefficient H^T for polytomously scored items.

2.3.2 Influence of Item Properties and Test Length on H^T

Figure 2.2 illustrates that for dichotomously scored items coefficient H^T cannot distinguish well between data generated under a model inconsistent with IIO (Figure 2.2a) and one consistent with IIO (Figure 2.2b). Here, the data was generated under the two-parameter logistic model with the same location

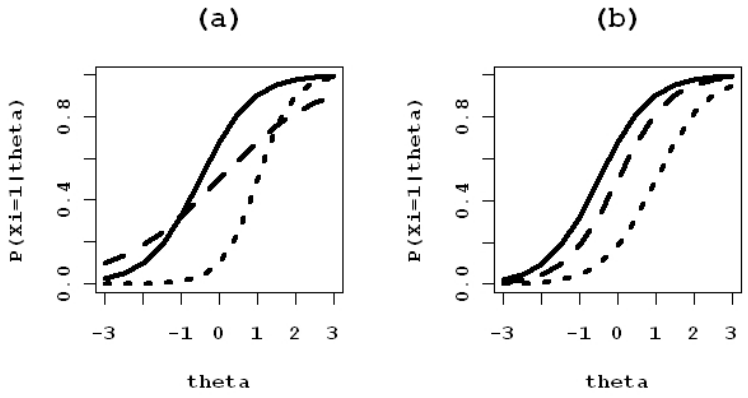

Figure 2.2: (a) Failure of IIO and (b) IIO. Both cases produce $H^T \approx .50$.

parameters and with equal mean discriminations[1]. Hence, Sijtsma and Meijer (1992) recommended using Mokken scale analysis (e.g., Sijtsma & Molenaar, 2002) to first identify and remove items that have flat IRFs and tend to produce many intersections with other, often steeper IRFs. For the remaining items, they suggested concluding that IIO held based on the value of H^T and the percentage of negative subject scalability values (not discussed here) did not exceed 10; else, IIO was rejected. Because we used method manifest IIO to select an item set that is consistent with IIO, the use of coefficient H^T sufficed.

In their Monte Carlo study, for dichotomous items Sijtsma and Meijer (1992) found that coefficient H^T increases as distance between item locations increases or item discrimination increases. here, we used a computational study for polytomously scored items based on stratified sampled values of θ (for which sample size did not play a role) to investigate IIO conditions so as to learn how H^T may be used once IIO has been ascertained by means of method manifest IIO. Similar to Sijtsma and Meijer (1992), we included in our design 1) Distance between Item Locations, 2) Item Discrimination, and 3) Test Length, but with more variation in levels. We expected similar trends influences on H^T as for dichotomously scored items. The factors 4) number of Score Categories and 5) Distance between adjacent ISRFs were unique for polytomous items.

Method

Coefficient H^T was computed at for the restricted rating scale version of the GRM (Equation 3), which implies IIO. The dependent variable was the expected value of coefficient H^T, for which the computational details are given

[1]For both Figure 2.2a and Figure 2.2b: $\beta_i = .5, 0, 1$, and $\theta \sim N(0, 1)$. In addition, for Figure 2.2a: $\alpha_i = 1.5, .75, 2.25$ and for Figure 2.2b: $\alpha_i = 1.5, 1.5, 1.5$.

in Appendix B. The levels of the five independent variables are listed below.

1. *Distance Item Locations.* Item locations λ_i were symmetrical relative to the mean of the distribution of θ ($\mu_\theta = 0$), and adjacent item locations were at a constant distance. The location of the least attractive item was chosen $\lambda_k = 0, 1$, and 2 (for $\lambda_k = 0$ all IRFs coincide), corresponding to a distance between the most and least attractive item of 0, 2, and 4, respectively. The distance between adjacent items depend further on Test Length k.

2. *Item Discrimination.* The discrimination parameters α_i ranged from weak discriminating to strong discriminating: .5, 1, 1.5, and 2.

3. *Test Length.* Test Length had three levels: $k = 5, 10$, and 15.

4. *Score Categories.* For numbers of Score categories were considered: $m + 1 = 2, 3, 5$, and 7, where $m + 1 = 2$ corresponds to dichotomously scored items.

5. *Distance ISRFs.* For dichotomously scored items, by definition $\epsilon_x = 0$, but for polytomously scored items, the parameters $\epsilon_1, \ldots, \epsilon_m$ may vary. Two variations were considered. First, the extremes were fixed $\epsilon_1 = -1$ and $\epsilon_m = 1$, with the other $m-2$ ISRFs located at equal distances between these extremes. Thus, for greater m the ISRFs were more densely located around the item location. Second, the distance between the locations of adjacent ISRFs was fixed at .5, which resulted in a greater dispersion of the ISRFs around the item location as m was greater.

Because dichotomously scored items only have one item step, the 2 cells in the design corresponding to Distance ISRFs collapsed, resulting in a design with $3 \times 4 \times 3 \times 4 \times 2 - 2 = 286$ cells.

Results

For the design factors typical of polytomous items, which are number of Score Categories and Distance between adjacent ISRFs, we found little effect on coefficient H^T (no more than a few hundredths between corresponding design cells). This justifies discussing results for only the simplest case of $m + 1 = 2$. For the cells concerning coinciding IRFs ($\lambda_k = 0$), we found that $H^T = 0$ (consistent with mathematical proof in Appendix A). Table 2.3 shows H^T-values for the combinations of Slope of the IRFs, Distance Item Locations ($\lambda_k = 1, 2$), and Test Length. Similar to results found by Sijtsma and Meijer (1992), Slope of the IRFs and Distance Item Locations had positive effects on H^T. Unlike their results, however, where the number of items had no significant negative effect for $k = 9$ and 18, our results show a negative effect of Test Length. This discrepancy can be explained by the levels we used for

Table 2.3: Values of Coefficient H^T for Levels of Slope of the IRFs, Distance Item Locations, and Slope of the IRFs.

Test Length	Distance Item Location	Slope of the IRFs			
		.5	1.5	2.5	3.5
5	2	.038	.145	.279	.413
	4	.135	.405	.610	.742
10	2	.031	.121	.239	.362
	4	.112	.355	.557	.698
15	2	.029	.114	.227	.346
	4	.106	.340	.540	.683

Test Length ($k = 5, 10$, and 15), where we found the largest decrease in H^T between $k = 5$ and 10. These results suggest that beyond approximately 10 items there is little to no effect of the number of items on the value of H^T.

Discussion

The computational results supported the expectation that when items are further apart, for a fixed θ the items' response probabilities show more variation and the ordering of a subject's item scores better resembles the ordering of the items' sum scores. Given IIO, coefficient H^T expresses the degree to which the ordering of the sum scores on the items is reflected by the individual vectors of item scores. The next section illustrates the practical use of method manifest IIO and the coefficient H^T.

2.4 Real-Data Example

Method manifest IIO and coefficient H^T were used for investigating whether IIO held in the two scales for measuring deference ($k = 9$) and achievement ($k = 10$) from the Dutch version of the Adjective Checklist (ACL; Gough & Heilbrun, 1980). The scales were not constructed with IIO in mind, but are well suited for demonstrating the exploratory use of method manifest IIO. Items consist of an adjective and five ordered score categories. Table 2.4 shows the item labels (negatively worded items were recoded). The subjects were $n = 433$ students, who were instructed to consider whether an adjective described their personality, and rate the answer category that fitted best to this description. Vorst (1992) collected the data, which are available from the R-package `mokken` (Van der Ark, 2007).

Prior to investigating IIO, following Sijtsma and Meijer (1992) a Mokken scale analysis was done on both subscales. Inclusion of all items resulted in

Table 2.4: Number of Violations for the Deference and Achievement Scale.

Deference		Achievement		
Items	Step	Items	Step	
	1		1	2
Impulsive	0	Quitting*	0	0
Demanding	0	Unambitious*	0	0
Forceful	0	Determined	0	0
Rebellious	0	Active	0	0
Uninhibited	0	Energetic	0	0
Bossy	0	Ambitious	1	0
Reckless	0	Alert	2	-
Boastful	0	Persevering	1	0
Conceited	0	Thorough	0	0
	0	Industrious	0	0
Coefficient H^T	.320			.116

* Negatively worded items.

$H = .307$ for scale Deference, and $H = .378$ for scale Achievement. Following Mokken and Lewis (1982), $.30 \leq H < .40$ stands for a weak scale.

For method manifest IIO, the IRFs were estimated after adjacent rest scores were joined until each group contained at least 86 respondents (Molenaar & Sijtsma, 2000, p. 67). The R package `mokken` (Van der Ark, 2007) was used for the computations. Table 2.4 shows for $minvi = .03 \times m = .12$ that scale Deference did not have significant violations of IIO, and that $H^T = .320$. The scale Achievement had two significant violations, both involving item Alert. Removal of this item resulted in a scale containing nine items for which IIO held. Coefficient H_g^T cannot be computed for subjects that have the same scores on all items. For the scale Achievement, six subjects were excluded for computing H^T as their H_g^T-values could not be computed. For the remaining 427 respondents we found $H^T = .116$. Support for IIO is stronger for Deference than for Achievement. Interpretation of H^T is discussed in the next section.

2.5 General Discussion

A top-down sequential procedure was used for selecting a subset of items having nonintersecting IRFs, based on method manifest IIO. Thus, not all item subsets were investigated and, once removed, an item was not re-evaluated for possible re-selection in later steps of the procedure. Alternative selection procedures (e.g., genetic algorithms; Michalewicz, 1996), which assess all possible item subsets, may be investigated in future research so that possibly larger and

different item subsets for which an IIO holds may be identified.

IIO research is new, and experience on how to interpret results has to accumulate as more applications become available. For the time being, we tentatively generalize the heuristic rules proposed by Mokken and Lewis (1982) for interpreting values of scalability coefficient H to the interpretation of H^T values, provided IIO holds for an item set. Thus, we propose: $H^T < .30$ means that the item ordering is too inaccurate to be useful; $.30 \leq H^T < .40$ means low accuracy; $.40 \leq H^T < .50$ means medium accuracy; and $H^T \geq .50$ means high accuracy. Based on these rules, the nine items from the Deference scale may be ordered with low accuracy ($H^T = .320$), and the remaining nine items from the Achievement scale do not have IIO ($H^T = .116$).

The assumption of IIO is both omnipresent and implicit in the application of many tests, questionnaires, and inventories. Test constructors and test users alike often assume that the same items are easy or attractive for each of the respondents to whom the items are administered but rarely put this strong assumption to the test of empirical evaluation. Yet, an established IIO underpins and greatly facilitates the interpretation of the test results, for example, when the test administration procedure is based on the ordering of the items from easiest to most difficult, the items reflect a developmental sequence of cognitive steps assumed to be the same for everyone, or when the set of items is assumed to reflect a hierarchical or cumulative structure. Invariant item ordering for polytomously scored items is an unexploited terrain. This study provides a first start for this topic, and shows directions for future explorations.

Appendix

A IIO implies $H_g^T \geq 0$ and $H^T \geq 0$

First, let θ_g and θ_h be the latent variable values of two arbitrary subjects g and h. IIO implies that for subject g and h the items are ordered in the same way; that is,

$$E(X_1|\theta_g) \geq E(X_2|\theta_g) \geq \cdots \geq E(X_k|\theta_g) \tag{A1}$$

and

$$E(X_1|\theta_h) \geq E(X_2|\theta_h) \geq \cdots \geq E(X_k|\theta_h). \tag{A2}$$

Second, for two arbitrarily selected items i and j let $E(X_i - X_j|\theta_g)$ and $E(X_i - X_j|\theta_h)$ denote the expected differences between the item scores for subjects g and h, respectively. Because for g and h all items are ordered in the same way (Equations A1 and A2), it follows that $E(X_i - X_j|\theta_g)$ and $E(X_i - X_j|\theta_h)$ have the same sign. Thus

$$E(X_i - X_j|\theta_g) \times E(X_i - X_j|\theta_h) \geq 0. \tag{A3}$$

Third, because of assumption LI $E(X_i - X_j|\theta_g)$ and $E(X_i - X_j|\theta_h)$ are independent. As a result, Equation A3 is equivalent to

$$E[(X_i - X_j|\theta_g)(X_i - X_j|\theta_h)] \geq 0. \tag{A4}$$

Fourth, the left-hand side of Equation A4 equals $2\text{cov}(\mathbf{X}_g, \mathbf{X}_h)$, so that

$$2\text{cov}(\mathbf{X}_g, \mathbf{X}_h) = E[(X_i - X_j|\theta_g)(X_i - X_j|\theta_h)] \geq 0. \tag{A5}$$

Fifth, coefficient H_g^T was defined as (Equation 6)

$$H_g^T = \frac{\sum\limits_{h \neq g} \text{cov}(\mathbf{X}_g, \mathbf{X}_h)}{\sum\limits_{h \neq g} \text{maxcov}(\mathbf{X}_g, \mathbf{X}_h)}. \tag{6}$$

By substituting Equation 6 into Equation 7, H^T can be rewritten as

$$H^T = \frac{\sum\limits_{g} \sum\limits_{h \neq g} \text{cov}(\mathbf{X}_g, \mathbf{X}_h)}{\sum\limits_{g} \sum\limits_{h \neq g} \text{maxcov}(\mathbf{X}_g, \mathbf{X}_h)}. \tag{A6}$$

It follows from Equation A5 that $\text{cov}(\mathbf{X}_g, \mathbf{X}_h) \geq 0$. Because $\text{maxcov}(\mathbf{X}_g, \mathbf{X}_h) \geq 0$ by definition, it follows that $H_g^T \geq 0$ and $H^T \geq 0$. Sixth, because $\text{cov}(\mathbf{X}_g, \mathbf{X}_h) \leq \text{maxcov}(\mathbf{X}_g, \mathbf{X}_h)$, it follows that $H^T \leq 1$ and $H_g^T \leq 1$.

When the k IRFs coincide, then $E(X_i - X_j|\theta_g) = E(X_i - X_j|\theta_h) = 0$ for all pairs of items. It follows from Equations A3, through A6, that then $H_g^T = H^T = 0$.

B Computational Procedure for H^T

Procedure for Determining Coefficient H_{ij} using the Weighted Number of Guttman Errors

Molenaar (1991) proposed an algorithm that uses weighted Guttman errors to compute Loevinger's (1948) coefficient H_{ij} for polytomously scored items. This algorithm is explained by means of the example in Table 2.5. Table 2.5 shows joint frequencies, marginal frequencies, and cumulative proportions of the scores of 178 respondents on items i and j. Let f_{x_i,x_j} denote the joint frequency of item-score pattern $(X_i = x_i, X_j = x_j)$, and let f_{x_i} denote the marginal frequency of $X_i = x_i$. The order of the cumulative proportions, $P(X_i \geq x)$ and $P(X_j \geq x)$, for $x = 1, 2$, and 3, determines the order of the item steps. In Table 2.5, this order is: $P(X_i \geq 1) = .984$, $P(X_i \geq 2) = .905$, $P(X_j \geq 1) = .854$, $P(X_j \geq 2) = .596$, $P(X_i \geq 3) = .517$, and $P(X_j \geq 3) = .163$. Thus, the order of the item steps from most attractive to least attractive

Table 2.5: Contingency Table for the Scores of 178 Respondents on Items i and j with Weights in Parenthesis

		Item j					
		$x = 0$	$x = 1$	$x = 2$	$x = 3$	f_{x_i}	$P(X_i \geq x)$
Item i	$x = 0$	**3** (0)	0 (2)	0 (4)	0 (7)	3	1.000
	$x = 1$	**4** (0)	7 (1)	3 (2)	0 (4)	14	0.984
	$x = 2$	**10** (0)	**22** (0)	**34** (0)	3 (1)	69	0.905
	$x = 3$	9 (2)	17 (1)	**40** (0)	**26** (0)	92	0.517
	f_{x_j}	26	46	77	29	178	
	$P(X_j \geq x)$	1.000	0.854	0.596	0.163		

Note: data from Weijmar, Schultz, and Van der Wiel (1991), as referred to in Molenaar (1991, p. 101); f_{x_i} is the marginal frequency of X_i, f_{x_j} is the marginal frequency of X_j.

is

$$X_i \geq 1, X_i \geq 2, X_j \geq 1, X_j \geq 2, X_i \geq 3, X_j \geq 3. \tag{B1}$$

It is assumed that in order to reach item-score pattern $(X_i = x_i, X_j = x_j)$, one starts at the most attractive step and then proceeds step-by-step until one arrives at $(X_i = x_i, X_j = x_j)$. It happens regularly that along the way one misses steps that are necessary to arrive at the pattern $(X_i = x_i, X_j = x_j)$. The indicator vector $\mathbf{r}_{x_i,x_j} = \left(r^{(1)}_{x_i,x_j}, \ldots, r^{(2m)}_{x_i,x_j}\right)$ (here, $2m = 6$) tracks the item steps that must be passed (score 1) and the item steps that are failed (score 0) until item-score pattern $(X_i = x_i, X_j = x_j)$ is reached. For example, in Table 2.5, to obtain item-score pattern $(X_i = 1, X_j = 2)$, the first, third, and fourth item step in Equation B1 must be passed; thus, $\mathbf{r}_{1,2} = (1, 0, 1, 1, 0, 0)$. To obtain the item-score pattern $(X_i = 0, X_j = 0)$, none of the item steps must be passed and $\mathbf{r}_{0,0} = (0, 0, 0, 0, 0, 0)$.

The $2m$ elements of indicator vector $\mathbf{r}_{x_i,x_j}$ form $m(2m-1)$ different pairs. A pair is discordant if the element pertaining to the less attractive item step equals 1 (i.e., the less attractive item step was passed) and the element pertaining to the more attractive item step equals 0 (i.e., the more attractive item step was missed). It may be verified that $\mathbf{r}_{1,2} = (1, 0, 1, 1, 0, 0)$ has 15 pairs; two pairs, $(r^{(2)}_{1,2}, r^{(3)}_{1,2})$ and $(r^{(2)}_{1,2}, r^{(4)}_{1,2})$, are discordant, the other 13 pairs are concordant.

An item-score pattern is called a *conformal pattern* if for each item step that was passed no previous item steps were missed. This means that for a conformal pattern, indicator vector $\mathbf{r}$ has no discordant pairs of elements, $(0, 1)$. In Table 2.5, the frequencies of the conformal patterns are shown in bold face. An item-score pattern is called a *Guttman error* if for at least one of the item steps that were passed, a previous item step was missed. This means that indicator vector $\mathbf{r}$ has at least one discordant pair of elements, $(0, 1)$. In

Table 2.5, the frequencies of the Guttman errors are shown in normal face.

Each item-score pattern $(X_i = x_i, X_j = x_j)$ is weighted by the total number of discordant pairs in $\mathbf{r}_{x_i,x_j}$, denoted w_{x_i,x_j}. Then, the total number of discordant pairs equals

$$w_{x_i,x_j} = \sum_{v=2}^{2m} r_{x_i,x_j}^{(v)} \sum_{u=1}^{v-1} \left(1 - r_{x_i,x_j}^{(u)}\right). \tag{B2}$$

In Equation B2, the term $\sum_{u=1}^{v-1} \left(1 - r_{x_i,x_j}^{(u)}\right)$ counts the zeroes in $\mathbf{r}_{x_i,x_j}$ before the vth entry. For the item-score pattern $(X_i = 1, X_j = 2)$, for which $\mathbf{r}_{1,2} = (1,0,1,1,0,0)$ and $\mathbf{1} - \mathbf{r}_{1,2} = (0,1,0,0,1,1)$, the corresponding weight equals

$$\begin{aligned} w_{1,2} &= \sum_{v=2}^{6} r_{1,2}^{(v)} \left(\sum_{u=1}^{v-1} \left[1 - r_{1,2}^{(u)}\right] \right) \\ &= 0(0) + 1(1+0) + 1(0+1+0) + 0(0+0+1+0) + \\ &\quad 0(1+0+0+1+0) \\ &= 0+1+1+0+0 = 2. \end{aligned}$$

In Table 2.5, the weights corresponding to all the score patterns are given in parenthesis. The conformal patterns have zero weight.

Next, let f_{x_i,x_j} and e_{x_i,x_j} indicate the frequency of observed score patterns and the frequency of expected score patterns in the group under marginal independence, respectively. The scalability of two polytomously scored items i and j can be computed by insertion of

$$F_{ij} = \sum_{x_i=0}^{m} \sum_{x_j=0}^{m} w_{x_i,x_j} f_{x_i,x_j}$$

and

$$E_{ij} = \sum_{x_i=0}^{m} \sum_{x_j=0}^{m} w_{x_i,x_j} e_{x_i,x_j}$$

in

$$H_{ij} = 1 - \frac{F_{ij}}{E_{ij}}. \tag{B3}$$

Computation of H^T

Latent trait θ was approximated using 50 discrete values $\theta_1, \ldots, \theta_{50}$, of which the values were chosen such that $P(Z > \theta_g) = \frac{g}{51}$, for $g = 1, \ldots, 50$. This yields: $\theta_1 \approx -2.06$, $\theta_2 \approx -1.76$, $\theta_3 \approx -1.56$, $\ldots$, $\theta_{50} \approx 2.06$. For given θ_g, the probability of obtaining score x on item i is derived from the ISRFs of the IRT

model (in this case Equation 3); that is

$$P(X_i = x|\theta_g) = P(X_i \geq x|\theta_g) - P(X_i \geq x + 1|\theta_g). \tag{B4}$$

For subject g with θ_g and subject h with θ_h, and one item i, the bivariate probability that $(X_{ig}, X_{ih}) = (x_g, x_h)$ is computed as the product of the corresponding marginal score probabilities obtained from Equation B4; that is,

$$P(X_{ig} = x_g, X_{ih} = x_h|\theta_g, \theta_h) = P(X_i = x_g|\theta_g)P(X_i = x_h|\theta_h).$$

Assume that the k items in the test are administered to these two respondents, and let f_{x_g,x_h} denote the number of times that they produce item-score pattern (x_g, x_h). Then, the expected value of f_{x_g,x_h} under the IRT model equals

$$E(f_{x_g,x_h}) = \sum_{i=1}^{k} P(X_i = x_g|\theta_g)P(X_i = x_h|\theta_h). \tag{B5}$$

Above, coefficient H_{ij} was computed on the basis of the values of f_{x_i,x_j}. Based on these values, an ordering of the item steps was established (Equation B1) and weights, w_{x_i,x_j}, were derived (Equation B2). Similarly, to compute H^T_{gh} the values of $E(f_{x_g,x_h})$ are tabulated, and based on these values an ordering of *subject steps* is established and weights, w_{x_g,x_h}, are derived. After f_{x_i,x_j} has been replaced by $E(f_{x_g,x_h})$, the computational procedure for H^T_{gh} is exactly the same as for H_{ij}, Hence, the expected number of weighted Guttman errors under the IRT model equals

$$F_{gh} = \sum_{x_g=0}^{m} \sum_{x_h=0}^{m} w_{x_g,x_h} E(f_{x_g,x_h}). \tag{B6}$$

Under the model of marginal independence, for a test containing k items the expected number of times that subject g has score x_g and subject h has score x_h equals

$$e_{x_g,x_h} = \frac{1}{k} \sum_{i=1}^{k} P(X_i = x_g|\theta_g) \sum_{i=1}^{k} P(X_i = x_h|\theta_h). \tag{B7}$$

Hence, under marginal independence the expected number of weighted Guttman errors equals

$$E_{gh} = \sum_{x_g=0}^{m} \sum_{x_h=0}^{m} w_{x_g,x_h} e_{x_g,x_h}. \tag{B8}$$

Under the IRT model coefficients H^T_{gh} and H^T are computed using F_{gh}

(Equation B6) and E_{gh} (Equation B8); that is,

$$H_{gh}^{T} = 1 - \frac{F_{gh}}{E_{gh}}, \tag{B9}$$

and

$$H^{T} = 1 - \frac{\sum\sum_{g<h} F_{gh}}{\sum\sum_{g<h} E_{gh}}. \tag{B10}$$

A Numerical Example

Consider two respondents g and h, with $\theta_g = -\frac{1}{3}$ and $\theta_h = \frac{3}{4}$, and two trichotomously scored items i and j with item parameters: $\alpha = \frac{3}{2}$, $\lambda_i = -1$, $\lambda_j = 1$, and $\epsilon_x = (-\frac{1}{2}, \frac{1}{2})$. Here we consider the GRM in Equation 3 for illustration. For the GRM we obtain

$$\begin{aligned}
&P(X_i \geq x|\theta_g) = (1.000,\ .852,\ .562),\\
&P(X_i \geq x|\theta_h) = (1.000,\ .967,\ .867),\\
&P(X_j \geq x|\theta_g) = (1.000,\ .223,\ .060),\ \text{and}\\
&P(X_j \geq x|\theta_h) = (1.000,\ .593,\ .245);
\end{aligned}$$

and applying Equation B4 yields

$$\begin{aligned}
&P(X_i = x|\theta_g) = (.148,\ .290,\ .562),\\
&P(X_i = x|\theta_h) = (.033,\ .100,\ .867),\\
&P(X_j = x|\theta_g) = (.777,\ .163,\ .060),\ \text{and}\\
&P(X_j = x|\theta_h) = (.407,\ .348,\ .245).
\end{aligned}$$

Next, applying Equation B5, we obtain

$$\begin{aligned}
E(f_{0,0}) &= \sum_{i=1}^{k=2} P(X_{gi} = 0, X_{hi} = 0|\theta_g, \theta_h)\\
&= .148 \times .033 + .777 \times .407 = .322\ .
\end{aligned}$$

For the nine item-score patterns the results are tabled below:

		h				
		0	1	2	$E(f_g)$	$P(X_g \geq x)$
	0	.322	.285	.319	.924	1.000
g	1	.076	.085	.291	.454	.538
	2	.043	.077	.502	.622	.311
$E(f_h)$		.440	.448	1.112	$k = 2$	
$P(X_h \geq x)$		1.000	.780	.556		

Note: $E(f)$ is the expected marginal frequency.

Because $P(X_h \geq 1) = .780$, $P(X_h \geq 2) = .556$, $P(X_g \geq 1) = .538$, and $P(X_g \geq 2) = .311$, the ordering of the subject-steps is

$$X_h \geq 1, X_h \geq 2, X_g \geq 1, X_g \geq 2. \tag{B11}$$

Constructing indictor vectors $\mathbf{r}_{x_h,x_h}$ based on the ordering in Equation B11, and applying Equation B2 to these indicator vectors the following weights for $E(f_{x_h,x_h})$ are obtained:

		h		
		0	1	2
	0	0	0	0
g	1	2	1	0
	2	4	2	0

For marginally independent respondents, the expected frequencies for item-score patterns are found by using Equation B7. This yields, for example,

$$\begin{aligned} e_{0,0} &= \tfrac{1}{k}\sum_{i=1}^{k=2} P(X_i = 0|\theta_g) \sum_{i=1}^{k=2} P(X_i = 0|\theta_h); \\ &= (.148 + .033) \times (.777 + .407) = .204 \ ; \end{aligned}$$

and for all nine score patterns we obtain

		h			
		0	1	2	$E(f_g)$
	0	.204	.207	.515	.924
g	1	.100	.101	.252	.454
	2	.137	.139	.346	.622
$E(f_h)$		.440	.448	1.112	2

Coefficient H^T_{gh} is obtained from Equation B6, Equation B8, and Equation B9, such that

$$\begin{aligned} F_{gh} &= 0(0.322) + 0(0.285) + 0(0.319) + 2(0.076) + 1(0.085) + \\ &\quad 0(0.291) + 4(0.043) + 2(0.077) + 0(0.502) \\ &= 0.563; \\ E_{gh} &= 0(0.204) + 0(0.207) + 0(0.515) + 2(0.100) + 1(0.101) + \\ &\quad 0(0.252) + 4(0.137) + 2(0.139) + 0(0.346) \\ &= 1.127; \text{ and} \\ H^T_{gh} &= 1 - F_{gh}/E_{gh} \\ &= 0.500. \end{aligned}$$

Finally, coefficient H^T can be obtained from Equation B10.

Chapter 3

Selection of Alzheimer Symptom Items with Manifest Monotonicity and Manifest Invariant Item Ordering*

Abstract

A procedure is proposed for selecting items from a test for which the assumptions of both manifest monotonicity and manifest invariant item ordering hold. The use of the procedure is illustrated by means of an application to data from an Alzheimer disease assessment†.

*This Chapter has been published as: Ligtvoet, R., Van der Ark, L. A., & Sijtsma, K. (2008). Selection of Alzheimer symptom items with manifest monotonicity and manifest invariant item ordering. In K. Shigemasu, A. Okada, T. Imaizumi, & T. Hoshino (Eds.), *New trends in psychometrics* (pp. 225-234). Tokyo: Universal Academic Press.

†The authors would like to thank Larry Hughes for providing the data.

3.1 Introduction

Nonparametric item response theory (IRT) is based on a minimal set of assumptions necessary to obtain useful measurement properties. A regularly used assumption is *local independence* (LI) of the item scores given the latent variables underlying these item scores. Often, IRT models assume only one latent variable, an assumption that agrees with the practical requirement that the test measures only one trait or ability. This assumption is known as *unidimensionality* (UD). Other assumptions concern the relationships between the items and this latent variable. One assumption is *monotonicity* (M; Junker, 1993) and another is *invariant item ordering* (IIO; Sijtsma & Junker, 1996). M means that the higher the score on the latent variable, the higher the expected item score. IRT models that assume LI, UD, and M allow ordinal measurement of subjects (Sijtsma & Molenaar, 2002). IIO means that the ordering of the items in a test according to their attractiveness is the same for all levels of the latent variable. Sijtsma and Junker (1996) discuss several practical testing situations in which latent IIO is important, such as intelligence testing and person-fit analysis.

This study provides the outline of a bottom-up procedure which assists the researcher in selecting from a larger set of items a subset of items for which both M and IIO hold. The proposed procedure is illustrated by an application to data from Alzheimer disease assessment at the Southern Illinois University School of Medicine.

3.2 Theory

We assume that LI and UD hold for the test under consideration. For k items indexed i ($i = 1, \ldots, k$; indices i' and j are also used), let X_i denote the item score, and let X_i have realizations $x_i \in \{0, \ldots, m_i\}$. For dichotomous scoring, $m_i = 1$, and for polytomous scoring $m_i \geq 2$. Let θ denote the latent variable. The conditional expected value $E(X_i|\theta)$ is known as the item response function (IRF).

3.2.1 Latent and manifest monotonicity

The assumption of M means that the IRFs are nondecreasing in θ (i.e., no strict increasingness is required); that is,

$$E(X_i|\theta) \text{ is nondecreasing in } \theta. \tag{M}$$

For dichotomously scored items for which the assumptions of LI, UD, and M hold, IRFs may be estimated from the test data as follows. Let Y be a conveniently chosen ordinal estimator of latent variable θ, and let us consider the IRF of item i. We define the total score on $k-1$ items in the test excluding

item i for which we seek to estimate the IRF, as $Y_i = \sum_{j \neq i} X_j$. It has been shown (Junker, 1993) that LI, UD, and M together imply that

$$E(X_i|Y_i = y) \text{ is nondecreasing in } y,$$

for $y = 0, \ldots, k-1$. This observable property, known as manifest M, can be estimated from the test data for each item. Manifest M does not imply M; this means that manifest M is a necessary condition for M. Thus, if manifest M holds in the data, this provides support but not proof of M, whereas deviations from manifest M are in conflict with M. Because the proof of manifest M does not use the number of items k, manifest M also holds for a two-item test containing items i and j, for which $Y = X_j$; that is, $E(X_i|X_j = 0) \leq E(X_i|X_j = 1)$. A violation of manifest M occurs each time we find that $E(X_i|X_j = 0) > E(X_i|X_j = 1)$.

Unfortunately, for polytomously scored items for which LI, UD, and M hold it has been shown that M does not imply manifest M (B. T. Hemker, in Junker & Sijtsma, 2000). This means that a sequence of expected values $E(X_i|Y_i = y)$, that is nondecreasing in y, need not support M, and a sequence that is not monotone need not be in conflict with M (this conclusion also holds when $Y = X_{i'}$). In practical data analysis, however, it seems reasonable to assume that little harm is done when researchers use such sequences heuristically for assessing M (Sijtsma & Meijer, 2007).

In this study, for polytomously scored items we use this heuristic strategy for expected item score X_i conditioned on only one item X_j, and employ manifest M as

$$E(X_i|X_j = x_j) \text{ is nondecreasing in } x_j, \tag{1}$$

for $x_j = 0, \ldots, m_j$. A violation of manifest M occurs each time we find for two item scores $0 \leq x_j < z_j \leq m_j$, that

$$E(X_i|X_j = x_j) > E(X_i|X_j = z_j). \tag{2}$$

Figure 3.1a shows an example of two monotone IRFs. Examples of violations are found in Figure 3.1b (solid curve; one violation) and Figure 1d (solid curve; two violations).

3.2.2 Latent and manifest invariant item ordering

IIO has been defined for polytomously scored items (Sijtsma & Hemker, 1998) with $m + 1$ score categories, and dichotomous scoring as a special case. This definition can be generalized to items from the same test having different numbers of score categories, by considering the conditional expectation of item i adjusted for the number of answer categories, $E(X_i|\theta)/m_i$. Let the attractiveness of item i be defined as the unconditional expectation, $E(X_i)/m_i$, and let the items be numbered such that $i < i'$ means that item i is less attractive

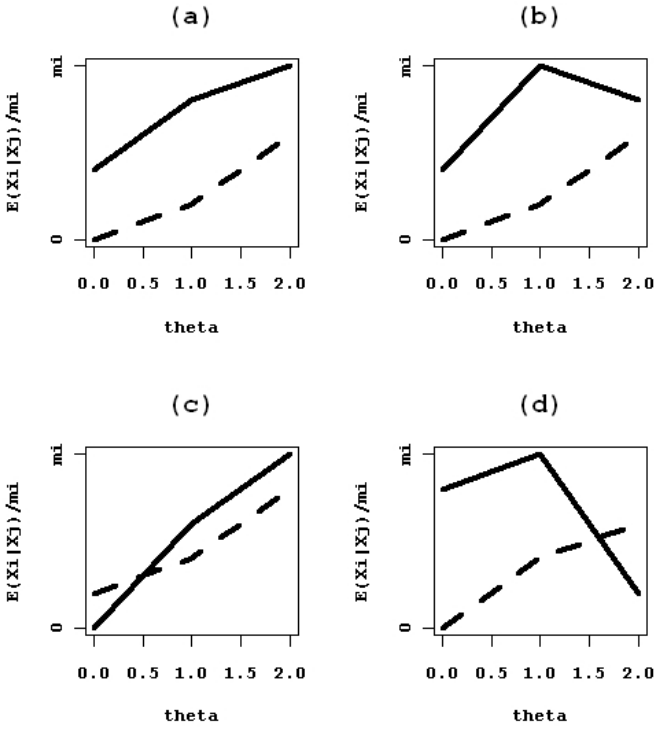

Figure 3.1: Graphs of $E(X_i|X_k)/m_i$ values with (a) no violations, (b) violation of manifest M, (c) violation of manifest IIO, and (d) violations of both manifest M and IIO.

than item i', $E(X_i)/m_i < E(X_{i'})/m_{i'}$. A set of k items has latent IIO if

$$E(X_i|\theta)/m_i \leq E(X_{i'}|\theta)/m_{i'}, \tag{IIO}$$

for all $i < i'$, and all θ. IIO allows possible ties. Ligtvoet, Van der Ark, Bergsma, and Sijtsma (2009) showed that for conditioning variable Y, item score X_i, and item score $X_{i'}$, all three independent of one another conditional on θ, latent IIO implies manifest IIO; that is

$$E(X_i|Y = y)/m_i \leq E(X_{i'}|Y = y)/m_{i'}, \tag{3}$$

for all $i < i'$, and all y; again allowing possible ties (proof in Appendix A). Manifest IIO does not imply latent IIO; thus, manifest IIO is a necessary condition for IIO. Analogous to manifest M, one may replace Y by a single item score, such that $Y = X_j$ and, as a result, Equation 3 is defined for a

3-tuple of items; that is,

$$E(X_i|X_j = x_j)/m_i \leq E(X_{i'}|X_j = x_j)/m_{i'}, \tag{4}$$

for all $i < i'$, and $x_j = 0, \ldots, m_j$. In this 3-tuple, each of the three items may play the role of conditioning variable Y. Thus, for a 3-tuple of items manifest IIO in Equation 4 needs to be evaluated three times, conditioned once on each of the three items i, i', and j. A violation of manifest IIO occurs each time we find for at least one value x_j, that

$$E(X_i|X_j = x_j)/m_i > E(X_{i'}|X_j = x_j)/m_{i'}. \tag{5}$$

Examples are found in Figure 3.1c (one violation at $x_j = 0$) and Figure 3.1d (one violation at $x_j = 2$).

3.2.3 Combination of monotonicity and invariant item ordering

For dichotomously scored items, M and IIO can be combined into one set of inequalities (Proposition 2.1 in Sijtsma & Junker, 1996). First, define the total score on $k-2$ items excluding the items i and i' for which we seek to establish non-intersection of the IRFs, such that $Y = Y_{ii'} = \sum_{j \neq i,i'} X_j$. By assuming that LI, UD, and M hold, and assuming that also IIO holds for items i and i', for two arbitrarily chosen values θ_a and θ_b, we have that if

$$\theta_a < \theta_b \Rightarrow E(X_i|\theta_a)/m_i \leq E(X_{i'}|\theta_b)/m_{i'}, \tag{6}$$

it follows that, for $Y = Y_{ii'}$,

$$y_a < y_b \Rightarrow E(X_i|Y_{ii'} = y_a)/m_i \leq E(X_{i'}|Y_{ii'} = y_b)/m_{i'}. \tag{7}$$

Equation 6 may be called assumption M&IIO, and Equation 7 may be called manifest M&IIO. For finite test length k, manifest M&IIO does not imply M&IIO. Manifest M&IIO is a necessary condition for M&IIO and, logically, if manifest M&IIO holds for the data this supports but does not prove M&IIO, whereas failure of manifest M&IIO disproves M&IIO. The proof that Equation 6 implies Equation 7 does not depend on the number of items k; thus, it also holds for $k = 3$ and $Y = Y_{ii'} = X_j$.

Because for polytomous scoring M does not imply manifest M, the implication in Equation 6 and Equation 7 does not straightforwardly generalize to polytomously scored items. Here we propose to use as a heuristic for investigating M&IIO (Equation 6) in real data the following manifest M&IIO property. Let the items i, i', and j be polytomously scored and let the number of ordered scores be variable across the items. Then, for two arbitrarily chosen item scores

for which $0 \leq x_j < z_j \leq m_j$, we propose to use the heuristic

$$E(X_i|X_j = x_j)/m_i \leq E(X_{i'}|X_j = z_j)/m_{i'}. \tag{8}$$

A violation of manifest M&IIO occurs each time we find for at least one pair $x_j < z_j$, that

$$E(X_i|X_j = x_j)/m_i > E(X_{i'}|X_j = z_j)/m_{i'}. \tag{9}$$

Figure 3.1b shows a violation of manifest M, which does not lead to a reversal of expectations as in Equation 9, Figure 3.1c shows a violation of manifest IIO, which again is not picked up, and Figure 3.1d shows a violation of both manifest M and manifest IIO, which is reflected by a reversal of expected values as in Equation 9. Thus, Equation 8 may not be a powerful tool for investigating violations of manifest M and manifest IIO.

3.2.4 Estimation of expected values

The $E(X_i|X_j)/m_i$ values for assessing the inequalities in the equations Equation 1, 4, and 8 can be estimated from the data as follows. Let n_{x_j} denote the number of respondents out of a sample of size n who have a score equal to x_j on item j, and let $n_{x_i x_j}$ denote the number of respondents who have a score x_i on item i and a score x_j on item j; then

$$\hat{E}(X_i|X_j = x_j)/m_i = \frac{1}{m_i} \sum_{x_i=1}^{m_i} x_i \frac{n_{x_i x_j}}{n_{x_j}}.$$

3.3 A bottom-up item selection procedure

The goal of this study is to suggest a procedure for finding subsets of items in a larger set, for which manifest M and manifest IIO are satisfied. The first step of the procedure is finding all 3-tuples consisting of three different items that satisfy manifest M and manifest IIO. The second step of the procedure entails combining the 3-tuples found in the first step into 4-tuples for which manifest M and manifest IIO hold. The third step entails combining the 4-tuples found in the second step into 5-tuples for which manifest M and manifest IIO hold. The procedure ends when the largest l-tuple is found for which manifest M and manifest IIO holds.

The first step of the procedure can be executed using two different methods for investigating manifest M and manifest IIO. The first method (*Method I*) investigates both manifest M (Equation 1) and manifest IIO (Equation 4). The second method (*Method II*) investigates manifest M&IIO (Equation 8). Method I and Method II are discussed next.

3.3.1 Technical details of first step

Method I. For the 3-tuple of items, i, i', and j, the procedure is as follows. Arbitrarily, let item j be the conditioning variable in Equation 1 and Equation 4. All violations of manifest M (Equation 2) for item i and item i' are tested using a one-sided independent t-test: The null hypothesis $E(X_i|X_j = x_j) \leq E(X_i|X_j = z_j)$ is tested against the alternative that $E(X_i|X_j = x_j) > E(X_i|X_j = z_j)$ (Equation 2). Rejection of the null hypothesis means that a violation of manifest M is found.

All violations of manifest IIO (Equation 5) are tested using a one-sided dependent t-test. For $E(X_i)/m_i < E(X_{i'})/m_{i'}$, the null hypothesis $E(X_i|X_j = x_j)/m_i \leq E(X_{i'}|X_j = x_j)/m_{i'}$ is tested against the alternative that $E(X_i|X_j = x_j)/m_i > E(X_{i'}|X_j = x_j)/m_{i'}$ (Equation 5). Rejection of the null hypothesis means that a violation of manifest IIO is found. It may be noted that testing may be problematic when X_i en $X_{i'}$ have not been measured on the same scale, but given the heuristic nature of this research this is ignored here.

Hypothesis testing is done with item i, i', and j consecutively playing the role of conditioning variable while the expected values of the other two items are evaluated. If no violations are found, this result supports M and IIO for the item 3-tuple. If at least one violation is found, it is concluded that M and IIO are not valid for the 3-tuple. Because each 3-tuple of items is investigated three times, and because the number of 3-tuples is large for realistic test length k, it is reasonable to expect large numbers of sample violations of manifest M and manifest IIO. Method I picks up all violations and allows the researcher to adjust the level of significance so that the violations considered to be important can be assessed for item selection with an eye to optimal decision-making.

Molenaar and Sijtsma (2000) noted that practical data analysis often yields large numbers of violations of manifest M and manifest IIO but argue that many of the relatively small violations are not damaging for the measurement of persons on an ordinal scale. They suggest ignoring violations smaller than a value *minvi* for statistical significance testing. The computer program MSP (Molenaar & Sijtsma, 2000) uses default option $minvi = .03$. If a large power of the statistical test is considered to be undesirable, the user could choose even larger values of *minvi*. For the violations that remain after selection by *minvi*, the nominal Type I error rate of the statistical test may be adapted. The requirement not to reject items too easily is accomplished using large *minvi* and small Type I error rate.

Another possibility arises in applications in which having the best item subset possible has priority over reliable subject ordering using large numbers of items. For example, when the importance of individual decision-making is paramount, as in medical diagnosis, items are selected that show no more than minor violations of manifest M and manifest IIO. This is accomplished using small *minvi* and large Type I error rate. In the research to be reported shortly both *minvi* and the nominal Type I error rate are manipulated.

Method II tests Equation 8, which is true when both manifest M and manifest IIO hold, using a dependent t-test. Again, $minvi = .03$ may be used for ignoring small sample violations, and statistical testing may be used for the remaining violations. However, the method may well overlook serious violations of manifest M and manifest IIO in the data and, as a result, it is expected to have less power than Method I.

3.3.2 Technical details of the next steps

The next steps combine item 3-tuples into larger items sets without further statistical testing. In the second step, item 4-tuples are identified of which all $\binom{4}{3} = 4$ constituent item 3-tuples were found in the first step by means of either Method I or Method II. In the third step, item 5-tuples are identified of which all $\binom{5}{4} = 5$ constituent item quartets were found in the second step; and so on, until no larger sets can be identified. The end result of this procedure may consist of several item subsets that overlap, and that may contain different numbers of items. It is up to the researcher to interpret this result with respect to his/her research question.

3.4 Simulation study

A study was done to investigate the effects of different values of *minvi* and the significance level α on the number of item 3-tuples identified using either Method I or Method II, and on the end result of item selection.

3.4.1 Method

The data were sampled from a data set consisting of the scores of 200 patients on the Mini-Mental State exam (MMS; Folstein, Folstein, & McHugh, 1975). These data were collected during an Alzheimer disease assessment at the Southern Illinois University School of Medicine between 1994 and 2000 (Hughes, Perkins, Wright, & Westrick, 2003). The MMS consists of eleven items that assess several cognitive functions such as orientation, registration, and attention (see Appendix B for item labels). The assumptions UD and Li were checked on the dichotomized item scores using the DETECT index (Zhang & Stout, 1999a, 1999b) and the scalability coefficient H (Mokken, 1971). The resulting values of DETECT and H suggested that the assumptions of UD and LI were satisfied by the data.

To investigate the effects of different test characteristics on the sets of items satisfying manifest M and manifest IIO, four design factors were used:

1. *Method* for investigating manifest M and manifest IIO. Method I and Method II were used.

2. *Size of minvi.* Two levels were investigated: $minvi = 0$, which implies that all observed violations are tested for significance, and $minvi = .03$, which is the MSP default value.

3. *Nominal Type I error rate.* The default value in MSP is $\alpha = .05$. In addition $\alpha = .10$ was considered, which leads to a more frequent rejection of the null hypothesis.

4. *Sample size.* A relatively small ($n = 200$) sample and a relatively large ($n = 500$) sample were drawn with replacement from the real data.

For each of the (Method $\times$ Size of *minvi* $\times$ Nominal Type I error rate $\times$ Sample Size) = 16 cells of the design, 10 data sets were generated. Given a particular Sample Size, a given data matrix was used across all 8 combinations of Method, Size of *minvi*, and Type I error rate. As a result, sample size is a between-factor, and the other independent variables are within-factors.

Two dependent variables were used in this study to assess the effect of the methods on the sets of items.

1. The proportion of item 3-tuples for which manifest M and manifest IIO could not be rejected. This proportion was computed as the number of item 3-tuples for which manifest M and manifest IIO could not be rejected divided by $\binom{11}{3} = 165$, the maximum number of item 3-tuples in this study.

2. The magnitude of the largest item set that resulted from the item selection procedure. The maximum value that can be obtained is 11.

3.4.2 Results

Table 3.1 shows the proportions and the standard deviations of the number of item 3-tuples in each cell of the design, based on the theoretical maximum of 165 item 3-tuples. Furthermore, Table 3.1 shows the modal number of items in the largest item set that resulted from the selection procedure. Most item 3-tuples were found using Method II; between 93% and 96% of the theoretical maximum. Here, the smaller standard deviation is due to the proportions being close to 1. This higher number of item 3-tuples leads to larger modal l-tuples; 8-tuples for $\alpha = .10$ and 9-tuples for $\alpha = .05$. The mean number of item 3-tuples was close to the theoretical maximum; thus, few 3-tuples violated M&IIO. For Method I, nominal Type I error rate and sample size had a relatively small effect. The modal number of items in the largest item set was equal to 6. Nominal Type I error $\alpha = .05$ yielded higher means than nominal Type I error $\alpha = .10$, and $n = 200$ yielded higher proportions than $n = 500$. These effects were smaller for Method II. No effect of factor *minvi* was found.

Table 3.1: Proportion of Item 3-tuples (Standard Deviation Between Parentheses) and Modal Number of Items in the Largest Set (bold).

	Size of		Sample Size	
Method	*minvi*	Alpha	200	500
I	.00	.05	.654 (.080) **6**	.623 (.076) **6**
		.10	.549 (.073) **6**	.530 (.072) **6**
	.03	.05	.654 (.080) **6**	.624 (.076) **6**
		.10	.552 (.075) **6**	.532 (.071) **6**
II	.00	.05	.957 (.011) **9**	.948 (.018) **9**
		.10	.939 (.017) **8**	.929 (.023) **8**
	.03	.05	.957 (.011) **9**	.948 (.018) **9**
		.10	.939 (.017) **8**	.929 (.023) **8**

3.4.3 Discussion

Method II identified few violations of M and IIO, and may have had too little power to be useful here. Lower bound *minvi* did not have any effect. This may be due to small sample size yielding no significance for small violations when $minvi = 0$. For $minvi = .03$, the same small violations were not tested or they were not significant, yielding the same result as found for $minvi = 0$. For larger sample sizes *minvi* is probably more effective for reducing power.

3.5 Real-data analysis

3.5.1 Method

Because the MMS is used to identify symptoms of Alzheimer disease, we considered reliably identifying the largest subset of items that is characterized by M and IIO to be of greatest importance here. Based on the results found in the previous section, we thus used Method I with $minvi = 0$ and $\alpha = .10$. These choices produced a conservative item selection, thus avoiding unnecessary risk of selecting items for which latent M and latent IIO did not hold.

3.5.2 Results

The 11 items from the MMS were numbered from least attractive to most attractive (see Appendix B for item numbers). Method I was used to test all 165 item 3-tuples three times for violations of manifest M and manifest IIO. This resulted in 107 item 3-tuples. Next, 3-tuples were combined to produce larger item sets fulfilling both manifest M and manifest IIO. The result was one 7-tuple containing the items 1, 2, 5, 6, 8, 9, and 11.

3.5.3 Discussion

The items 1, 2, 5, 6, 8, 9, and 11 can be interpreted as follows with respect to M and IIO. Given the severity of the symptoms associated with the items, we conclude that the first problems of Alzheimer disease occur with recalling recently learned series (item 11), naming the date (item 9), naming a location (item 8), following a sequence of instructions (item 6), copying a drawing (item 5), registration of words read aloud (item 2), and writing a sentence read aloud (item 1). This ordering seems to agree with the onset of symptoms (memory impairment, disorientation, impaired judgement, and language disturbance) reported by (Kang, Jeong, Lee, Baek, Kwon, Chin & Na, 2004). Much research with respect to the progression of Alzheimer disease remains to be done (and is beyond our expertise).

3.6 General discussion

A bottom-up item selection procedure was proposed which assists the researcher in selecting items from a larger set in which items satisfy the requirements of M and IIO. Two methods were used to assess these properties. Method I, which assesses both manifest M (Equation 1) and manifest IIO (Equation 4) is more demanding than Method II, which assesses a weaker version of manifest M and manifest IIO, here denoted *manifest M&IIO* (Equation 8). Any subset of items that is selected using Method I is also selected using Method II but not the other way round. Method II probably will gain power if it is used to assess expected values conditional on, for example, $Y = Y_{ii'}$, as in Equation 7. Rest score $Y = Y_{ii'}$ is certainly a more fine-grained ordinal estimator of latent trait θ than the coarse estimator provided by single item score $Y = X_j$ (Equation 8). Thus, rest score $Y = Y_{ii'}$ will reveal violations of *manifest M&IIO* more easily than item score $Y = X_j$. Use of the rest score requires the number of item scores to be the same across the items; else, scores from different items are incomparable.

For the analysis of the MMS we chose $minvi = 0$ and $\alpha = .10$. This way, we reduced the risk of selecting items for which manifest M and/or manifest IIO do not hold. The resulting item subset has these properties with much certainty; thus, we infer M and IIO to hold for these items. In other applications, in which a large number of items is needed to accurately order respondents for whom the items have the same ordering, a larger value of $minvi$ and a smaller value of α allow more items into the scale but the selection is less stringent and accepts more violations of M and IIO. Future research should clarify which values of $minvi$ and α are acceptable for producing scales that allow for accurate subject and item ordering, while rejecting as few items as possible.

This study did not provide a benchmark indicating whether Method I or Method II is more appropriate for identifying violations of M and IIO. Such a

study would require knowledge about the presence of M and IO in the population, but this information is unavailable in real-data analysis. A well controlled simulation study to investigate the extend to which the procedure yields correct conclusions concerning M and IIO, is a topic for future research.

Furthermore, it has been suggested to use scalability coefficient H to evaluate manifest M (Molenaar, 1991; Mokken, 1971, pp. 148–153) and coefficient H^T to evaluate manifest IIO (Ligtvoet, Van der Ark, Te Marvelde, & Sijtsma, in press; Sijtsma & Meijer, 1992). For the item subset 1, 2, 5, 6, 8, 9, and 11, we found that $H = .598$, which indicates a "strong" scale (Mokken, 1971, p. 185). This result supports manifest M. We also found that $H^T = .747$, which indicates a strong agreement of subjects' item-score patterns with the ordering of items in the group. This result supports manifest IIO. These results lend credibility to the results obtained by means of our proposed item selection procedure.

Appendix

A IIO implies manifest IIO

The proof that IIO implies manifest IIO (Equation 3) is based on a proof by Van der Ark and Bergsma (2006). Let $G(\theta)$ be the distribution function of θ.

First, $E(X_i|Y)/m_i$ (see Equation 3) is rewritten. Standard algebra shows that

$$\begin{aligned} E(X_i|Y)/m_i &= \frac{1}{m_i}\sum_{x_i} x_i P(X_i = x_i|Y) = [m_i \cdot P(Y)]^{-1}\sum_{x_i} x_i P(X_i = x_i, Y) \\ &= [m_i \cdot P(Y)]^{-1}\int_\theta \sum_{x_i} x_i P(X_i = x_i, Y, \theta) dG(\theta) \\ &= [m_i \cdot P(Y)]^{-1}\int_\theta \sum_{x_i} x_i P(X_i = x_i, Y|\theta) G(\theta) dG(\theta). \end{aligned} \tag{A1}$$

Because of LI, $P(X_i = x_i, Y|\theta)$ in Equation A1 can be further reduced to

$$P(X_i = x_i, Y|\theta) = P(X_i = x_i|\theta)P(Y|\theta). \tag{A2}$$

Substituting Equation A2 into Equation A1 gives

$$\begin{aligned} E(X_i|Y)/m_i &= [m_i \cdot P(Y)]^{-1}\int_\theta P(Y|\theta)G(\theta)\sum_{x_i} x_i P(X_i = x_i|\theta) dG(\theta) \\ &= \frac{1}{m_i}\int_\theta E(X_i|\theta)G(\theta|Y) dG(\theta). \end{aligned} \tag{A3}$$

Second, it is shown that IIO implies a manifest IIO. Multiplying both sides

of the definition of IIO by $G(\theta|Y)$ and taking the integral over $G(\theta)$ leaves the inequality unchanged. Hence, IIO implies

$$\frac{1}{m_i}\int_\theta E(X_i|\theta)G(\theta|Y)dG(\theta) \leq \frac{1}{m_j}\int_\theta E(X_j|\theta)G(\theta|Y)dG(\theta). \qquad \text{(A4)}$$

It follows from Equation A3 that the left-hand side of Equation A4 equals $E(X_i|Y)/m_i$, and the right-hand side of Equation A4 equals $E(X_j|Y)/m_j$. Hence, IIO implies Equation 3. This completes the proof.

B The 11 Alzheimer assessment items.

For the 11 Alzheimer assessment items from the Mini-Mental State exam, Appendix B shows the item numbers, their content, and their number of answer categories between parentheses.

1. Writing (2)
2. Registration (2)
3. Repetition (2)
4. Reading (2)
5. Copying (2)
6. 3-Stage Command (3)
7. Attention (4)
8. Location (4)
9. Date (5)
10. Naming (2)
11. Recall (4)

Chapter 4

Latent Class Models for Testing Monotonicity and Invariant Item Ordering for Polytomous Items*

Abstract

Two assumptions that are relevant to many applications using item response theory (IRT) are the assumptions of monotonicity (M) and invariant item ordering (IIO). An IRT model is proposed for ordinal items which implies M and IIO. This model is specified as a latent class model with inequality constraints on the class-specific item means. A Gibbs sampling scheme is used for estimating the model parameters. It is shown that the deviance information criterium (DIC) can be used as an overall test of M and IIO, while posterior predictive checks can be used to test these assumptions at the item level. A real-data application illustrates a model fitting strategy for selecting an item set for which M and IIO holds.

*This Chapter has been submitted for publication.

4.1 Introduction

Item response theory (IRT) models are used to construct measures using multiple observed scores from tests or questionnaires. An example of a questionnaire item used to measure a subject's attitude towards women's liberation reads "*Women's liberation sets women against men.*", for which the subject scores 0 if he or she *agrees* with the statement, 1 if he or she is *neutral* towards the statement, and 2 if he or she *disagrees* with the statement (Heinen, 1996, p. 291). The score on each item i is considered to be a random variable X_i, for which each subject has the ordered score $x_i \in \{0, \ldots, m\}$, and where for the example $m = 2$. In IRT an unobservable *latent variable* θ is postulated to underly the items scores and the item scores are assumed to be mutually independent conditional on θ. The latter assumption is usually referred to as the assumption of *local independence* (Ip, 2001; Lord & Novick 1968, p. 361). Further, in IRT the relationship between the latent variable and the item scores is described by means of response functions, and it are these response functions about which specific assumptions are made. Let a_i be the slope parameter of item i and b_{x_i} its difficulty parameter corresponding to item score $X_i = x_i$. Samejima's (1969, 1997) graded response model, for example, specifies a logistic function for the cumulative probability of item i ($i = 1, \ldots, k$) and score $x_i = 1, \ldots, m$; that is,

$$\text{logit}P(X_i \geq x|\theta) = a_i(\theta - b_{x_i}). \tag{1}$$

with $a_i > 0$. In this IRT model, the $P(X_i \geq x|\theta)$ are nondecreasing in θ, which implies that subjects with higher scores on θ (e.g., with a more extreme attitude related to women's liberation) are expected to score higher on each of the items. The nondecreasingness of the response functions is referred to as the assumption of *monotonicity* (M; e.g., Holland & Rosenbaum, 1986).

The logistic shape of the response function in Equation 1 is mathematically convenient, but it may also be too restrictive, in which case model-data misfit may occur while M may still hold. When solely interested in the assumption M, more flexible models may be considered which still allow for an ordering of the subjects on θ (e.g., Junker & Sijtsma, 2001; Molenaar, 1997). These subject ordering models are based on the assumption M, which states that for all i,

$$E(X_i|\theta), \text{ is nondecreasing in } \theta. \tag{M}$$

In contrast to most IRT models which define m response functions for each item, taking the conditional expectation $E(X_i|\theta)$ as the point of departure yields an IRT model with only one response function per item.

In addition to assumption M relating to the ordering of subjects based on θ, a second assumption relating to the ordering of the items is sometimes considered. The latter assumption, which is known as an *invariant item ordering* (IIO; Sijtsma & Hemker, 1998), implies the same ordering of items, in terms of

attractiveness (or difficulty), holds for all subjects. Let the items be indexed so that $i < i'$ indicates that $E(X_i) \leq E(X_{i'})$, then IIO states that

$$E(X_1|\theta) \leq E(X_2|\theta) \leq \cdots \leq E(X_k|\theta), \text{ for all } \theta. \tag{IIO}$$

Sijtsma and Junker (1996) provided various examples of practical applications of IRT requiring IIO. However, only few IRT models imply IIO (Sijtsma & Hemker, 1998), and those models that do so are often too restrictive to fit real life data well. For example, while the IRT model described in Equation 1 does not imply IIO, a version with equal slope parameters across items and difficulty parameters restricted by $b_{x_i} = d_i + c_x$ yields a model that is consistent with the IIO assumption. The resulting rating scale version of the grade response model (Sijtsma & Hemker, 1998) is defined as

$$\text{logit}P(X_i \geq x|\theta) = a(\theta - d_i - c_x). \tag{2}$$

This parametric IRT model will typically be too restrictive to show an acceptable model-data fit, whereas in reality the M and IIO assumptions may still hold. The reason is that it imposes many more constraints than M and IIO. The contribution of this article is that it proposes a class of less restricted IRT models that in addition to assumption M can be used to test assumption IIO without imposing "irrelevant" restrictions such as logistic response functions. The resulting non-parametric IRT models can be considered additions to existing tools for IIO research (see also, Ligtvoet, Van der Ark, Te Marvelde, & Sijtsma, in press).

Constrained latent class models (LCMs) have been shown to be useful tools for obtaining approximations of IRT models, where the underlying continuous latent variable is discretized by assuming that subjects belong to q homogenous latent classes (Heinen, 1996). By imposing linear constraints on the logistic parameters of a LCM, discretized versions of parametric IRT models, such as the graded response model, can be obtained (Vermunt, 2001). The approach proposed in this paper is to assess the assumptions M and IIO using latent class models with inequality restrictions. Such LCMs with inequality constraints on the model parameters have been used for testing model assumptions in the context of nonparametric IRT models (e.g., Croon, 1990, 1991; Karabatsos & Sheu, 2004; Vermunt, 2001). Also Hoijtink and Molenaar (1997; see also Hoijtink, 1998; Van Onna, 2002) showed how to formulate nonparametric IRT models as LCMs with ordinal constraints, and moreover illustrated how to estimate and test such models using Bayesian methods. Here, we apply a similar type of Bayesian procedure to formulate a non-parametric IRT model for the conditional expected item scores to test assumption M and IIO. As an alternative, likelihood based methods developed by Bartolucci and Forcina (2005), and Vermunt (1999, 2001) could be adapted to the models of interest in this paper, but this approach is not further pursuit here.

The remainder of this paper is organized as follows. First, we present the unrestricted LCM and explain how to estimate its parameters using a Gibbs sampler. Then we describe the restricted LCMs corresponding to M and IIO and show how to incorporate the implied inequality restrictions in a Gibbs sampler. Subsequently, Bayesian tests for assessing the validity of the M and IIO assumption are described. And finally, the proposed LCM procedure for testing M and IIO is illustrated with an application using five questionnaire items measuring respondents' attitudes toward women's liberation.

4.2 Unrestricted Latent Class Models

To test the assumptions M and IIO, we approximate the continuous latent variable in IRT using a finite number of latent classes (i.e., $\theta = 1, \ldots, q$). As an IRT model, the resulting latent class model is a model for $P(\mathbf{x})$; that is, a model for a particular response pattern (note that $\mathbf{x}$ refers to the vector of k item scores). LCMs are finite mixture models (Agresti, 1990; McLachlan & Peel, 2000), where subjects belonging the same latent class are homogeneous with respect to the probability $P(\mathbf{x}|\theta)$ of obtaining a certain response pattern. In addition, we assume that within each latent class, the item scores are mutually independent (e.g., Lazarsfeld & Henry, 1968, p. 22), which is equivalent to the IRT assumption of local independence (see Clogg, 1988). Let $\pi_{x_i|\theta}$ denote the probability of a score $X_i = x_i$ given θ and letting π_θ be the class proportion, the probability $P(\mathbf{x})$ is expressed as

$$P(\mathbf{x}) = \sum_\theta \pi_\theta P(\mathbf{x}|\theta) = \sum_\theta \pi_\theta \prod_i \pi_{x_i|\theta}. \tag{3}$$

Apart from choosing a fixed number of latent classes q and the local independence assumption, Equation 3 is yet unconstrained and referred to as the unconstrained LCM.

The Gibbs sampler is an iterative algorithm that can be used for obtaining samples from the posterior distribution of the parameters of a statistical model given a set of data, a likelihood function linking the data to the model of interest, and a prior distribution for the unknown parameters (e.g., Zeger & Karim, 1991). Let t denote the tth iteration of the Gibbs sampler algorithm. When applied to the model described in Equation 3, the algorithm starts by assigning initial values the class proportions

$$\pi_\theta{}^{(0)} = \left(\pi_1^{(0)}, \ldots, \pi_q^{(0)}\right)$$

and the conditional item probabilities

$$\begin{aligned}\pi_{\mathbf{x}}{}^{(0)} \quad = \quad & \big(\pi^{(0)}_{0_1|1}, \ldots, \pi^{(0)}_{m_1|1}, \cdots, \pi^{(0)}_{0_k|1}, \ldots, \pi^{(0)}_{m_k|1}, \\ & \cdots, \pi^{(0)}_{0_1|q}, \ldots, \pi^{(0)}_{m_1|q}, \cdots, \pi^{(0)}_{0_k|q}, \ldots, \pi^{(0)}_{m_k|q}\big).\end{aligned}$$

The subsequential three steps are passed iteratively. The first step is a data augmentation step, which involves assigning each subject to one of the latent classes (Tanner & Wong, 1987). The second and third steps consist of sampling values for the class proportions $\pi_\theta{}^{(t)}$ and the conditional item probabilities $\pi_{\mathbf{x}}{}^{(t)}$, respectively. The algorithm is repeated until a criterium of convergence is reached. Ones this criterium is reached, the parameters are sampled in successive iterations as if sampled from their posterior distribution (Gelfand, Smith, & Lee, 1992). Based on these samples from the posterior, inferences can then be made about the parameter values. Our Gibbs sampler algorithm consists of the following three steps:

Step 1: Given the parameters values at the previous iteration, we derive from Equation 3 for each subject j with a score pattern $\mathbf{x}$ the probability of belonging to latent class θ as

$$P\big(\theta^{(t)}|\mathbf{x}_j\big) = \frac{\pi_\theta^{(t-1)}}{C} \prod_i \pi^{(t-1)}_{x_i|\theta},$$

with

$$C = \sum_{\theta=1}^{q} \pi_\theta^{(t-1)} \prod_i \pi^{(t-1)}_{x_i|\theta}.$$

Each subject is assigned to a latent class by a single draw from a multinomial distribution. This augmentation step yields values of the number of subjects in latent class θ, denoted $n_\theta^{(t)}$, and values of the number of subjects in latent class θ with the item score x_i on item i, denoted $n^{(t)}_{x_i|\theta}$. The $n_\theta^{(t)}$ and $n^{(t)}_{x_i|\theta}$ are used in the next two steps of the algorithm.

Step 2: The number of subjects the latent class θ ($n_\theta^{(t)}$) can be assumed to be defined by a multinomial distribution, which combined with a (conjugate) Dirichlet prior distribution yields a Dirichlet posterior distribution for $\pi_\theta{}^{(t)}$:

$$\pi_\theta{}^{(t)} \sim \mathrm{Dir}(n_1^{(t)} + \alpha, \ldots, n_q^{(t)} + \alpha).$$

Here, the α are hyper-parameters which are chosen to equal unity; reflecting ignorance concerning the information of the prior.

Step 3: Likewise, for a given item i and θ the parameters $\pi^{(t)}_{0_i|\theta}, \ldots, \pi^{(t)}_{m_i|\theta}$ are

sampled form a Dirichlet distribution

$$\pi^{(t)}_{0_i|\theta}, \ldots, \pi^{(t)}_{m_i|\theta} \sim \text{Dir}(n^{(t)}_{0_i|\theta} + \alpha, \ldots, n^{(t)}_{m_i|\theta} + \alpha).$$

4.3 Latent Class Models with M and IIO Constraints

Here, we wish to test whether the assumptions M and IIO hold by making use of LCMs. Recall that these assumptions imply particular inequality constraints on $E(X_i|\theta)$ across θ values and across items, respectively. Because $E(X_i|\theta) = \sum_x x_i \pi_{x_i|\theta}$ and $\pi_{x_i|\theta}$ are the LCM parameters, imposing the M and IIO constraints in a LCM implies that the inequality constraints on $E(X_i|\theta)$ should be translated into inequality constraints on the $\pi_{x_i|\theta}$. Once this is done, the restricted $\pi_{x_i|\theta}$ can be estimated using Gibbs sampling from truncated Dirichlet distributions; that is, from distributions in which the $\pi_{x_i|\theta}$ parameters can only attain values which are in agreement with M and IIO. However, the sampling of $\pi_{x_i|\theta}$ under the relevant constraints is complicated by the fact that the $m+1$ probabilities for item i and class θ are not independent of one another. Van Onna (2002) proposed resolving this issue by a sequential sampling scheme in which the probabilities for categories 0 to $m-1$ are sampled from truncated Beta distributions and the probability for category m is obtained by $\pi^{(t)}_{m_i|\theta} = 1 - \sum_{x=0}^{m-1} \pi^{(t)}_{x_i|\theta}$. Because a small simulation (not reported here) revealed that this procedure does not yield correct samples from the posterior distribution, instead we follow Hoijtink (1998, see also Laudy & Hoijtink, 2007) suggestion and re-parameterize $\pi_{x_i|\theta}$ as

$$\pi_{x_i|\theta} = \frac{\gamma_{x_i|\theta}}{\sum\limits_x \gamma_{x_i|\theta}}, \tag{4}$$

where $\gamma_{x_i|\theta} \sim \text{Gamma}(n_{x_i|\theta} + \alpha, 1)$. It is important to note that also $E(X_i|\theta)$ can be expressed in terms of these $\gamma_{x_i|\theta}$ parameters; that is,

$$E(X_i|\theta) = \frac{\sum\limits_x x\gamma_{x_i|\theta}}{\sum\limits_x \gamma_{x_i|\theta}}. \tag{5}$$

The advantage of this re-parametrization is that in contrast to the $\pi_{x_i|\theta}$, the $\gamma_{x_i|\theta}$ can be sampled independently of one another (e.g., Ferguson, 1973; Narayanan, 1990). It is well-known that sampling $m+1$ parameters $\gamma_{x_i|\theta}$ from independent Gamma distributions is equivalent to sampling the $m+1$ parameters $\pi_{x_i|\theta}$ from a $(m+1)$-dimensional Dirichlet distribution.

4.3.1 Monotonicity Constraints

Implementation of the assumption M in a LCM means that in addition to the model formulated in Equation 3, for each item i, $E(X_i|\theta-1) \leq E(X_i|\theta)$ for $\theta = 2,\ldots,q$ and $E(X_i|\theta) \leq E(X_i|\theta+1)$ for $\theta = 1,\ldots,q-1$. Substitution of $E(X_i|\theta)$ by its definition in Equation 5 yields

$$E(X_i|\theta-1) \leq \frac{\sum\limits_x x\gamma_{x_i|\theta}}{\sum\limits_x \gamma_{x_i|\theta}} \tag{6}$$

$$\frac{\sum\limits_x x\gamma_{x_i|\theta}}{\sum\limits_x \gamma_{x_i|\theta}} \leq E(X_i|\theta+1). \tag{7}$$

for $2 \leq \theta \leq q$ and $1 \leq \theta \leq q-1$, respectively.

In the Gibbs sampler, $\gamma_{x_i|\theta}$ is sampled at each iteration given the values of all the other parameters. This means that the restrictions on $\gamma_{x_i|\theta}$ implied by M should be derived from Equations 6 and 7; that is, the bounds for the admissible values of $\gamma_{x_i|\theta}$ are obtained by isolating the term $\gamma_{x_i|\theta}$ from these equations. More specifically, we can derive the first bound for $\gamma_{x_i|\theta}$ by using the equality $E(X_i|\theta-1) = E(X_i|\theta)$. Denoting this bound by u^{M}, we obtain

$$u^{\mathrm{M}}_{x_i|\theta} = \frac{\sum\limits_{y\neq x} \gamma_{y_i|\theta}\Big(E(X_i|\theta-1) - y\Big)}{x - E(X_i|\theta-1)},$$

for $\theta \geq 2$. The second bound corresponds to the value of $\gamma_{x_i|\theta}$ for which $E(X_i|\theta) = E(X_i|\theta+1)$. Denoting this bound by v^{M}, for $\theta < q$

$$v^{\mathrm{M}}_{x_i|\theta} = \frac{\sum\limits_{y\neq x} \gamma_{y_i|\theta}\Big(E(X_i|\theta+1) - y\Big)}{x - E(X_i|\theta+1)}.$$

Though at first glance one may think that $u^{\mathrm{M}}_{x_i|\theta}$ and $v^{\mathrm{M}}_{x_i|\theta}$ are the lower and upper bounds for $\gamma_{x_i|\theta}$, respectively, this is not correct. To illustrate this point, consider $\gamma_{0_i|1}$, the parameter for the first category of item i at $\theta = 1$. Its bound is derived from the equality $E(X_i|\theta=1) = E(X_i|\theta=2)$. Here, v^{M} is not an upper but a lower bound for $\gamma_{0_i|1}$, because gamma values smaller than v^{M} yield higher $E(X_i|\theta=1)$ values, which are not allowed according to the M restriction. It turns out that the bounds define the domain of admissible values for $\gamma_{x_i|\theta}$, which can lie either inside or outside the interval defined by $u^{\mathrm{M}}_{x_i|\theta}$ and $v^{\mathrm{M}}_{x_i|\theta}$. This means that not only the bounds should be computed at each step of the Gibbs sampler, but it should also be checked whether the domain of admissible values lies inside or outside the bounds. Denoting the admissible

domain under assumption M by $\mathcal{A}^{\mathrm{M}}_{x_i|\theta}$, Step 3 of the Gibbs sampler can be implemented as follows:

Step 3*: For each $\gamma_{x_i|\theta}$ at each iteration t compute the bounds $u_{x_i|\theta}$ and $v_{x_i|\theta}$ and sample a new value $\gamma^{(t)}_{x_i|\theta}$ from a truncated Gamma distribution:

$$\gamma^{(t)}_{x_i|\theta} \sim \text{Gamma}(n^{(t)}_{x_i|\theta} + \alpha, 1 \big| \gamma^{(t)}_{x_i|\theta} \in \mathcal{A}_{x_i|\theta}).$$

After each sample of $\gamma^{(t)}_{x_i|\theta}$, $E(x|\theta)^{(t)}$ is re-computed using Equation 5.

The method of inverse probability sampling is used to obtain samples from the relevant truncated gamma distributions (e.g., Gelfand, Smith, & Lee, 1992).

4.3.2 Invariant Item Ordering Constraints

The IIO constraints on the LCM are similar to the ones for M, but with the role of the items and latent classes reversed; that is, for any class q, $E(X_{i-1}|\theta) \leq E(X_i|\theta)$ for $i = 2, \ldots, k$ and $E(X_i|\theta) \leq E(X_{i+1}|\theta)$ for $i = 1, \ldots, k-1$. The bounds on $\gamma_{x_i|\theta}$ under IIO are

$$u^{\mathrm{IIO}}_{x_i|\theta} = \frac{\sum\limits_{y \neq x} \gamma_{y_i|\theta} \Big(E(X_{i-1}|\theta) - y \Big)}{x - E(X_{i-1}|\theta)},$$

for $i \geq 2$, and

$$v^{\mathrm{IIO}}_{x_i|\theta} = \frac{\sum\limits_{y \neq x} \gamma_{y_i|\theta} \Big(E(X_{i+1}|\theta) - y \Big)}{x - E(X_{i+1}|\theta)},$$

for $i < k$. At each iteration of the Gibbs sampler, both bounds are computed and the admissible domain $\mathcal{A}^{\mathrm{IIO}}_{x_i|\theta}$ is determined. To constrain the LCM by both M and IIO, the $\gamma^{(t)}_{x_i|\theta}$ parameters are sampled from truncated Gamma distributions with admissible ranges $\mathcal{A}_{x_i|\theta}$ defined as the intersection of $\mathcal{A}^{\mathrm{M}}_{x_i|\theta}$ and $\mathcal{A}^{\mathrm{IIO}}_{x_i|\theta}$.

4.3.3 Assessing Convergence

The values of the parameters obtained from the Gibbs sampler can be considered samples from their posterior distribution as the number of iterations t of the Gibbs sampler approaches infinity. In practice however, we are interested in t to be sufficiently large for our samples to correctly approximate the posterior distributions. The first samples of the model parameters are drawn after discarding the initial 10000 parameter values corresponding to the burn-in period. Sequential samples are drawn at intervals of 10 iterations. For these samples,

convergence is first assessed by comparing the samples from the posteriors between successive samples of size 5000 (e.g., Hoijtink & Molenaar, 1997). If the differences between the posteriors are small, it is concluded that convergence is reached. If the differences are large, the samples are discarded and new samples are taken until it can be concluded that convergence is reached. A second assessment for convergence we used is by inspection of the trace lines for the likelihood function across iterations of the Gibbs sampler (e.g., Gamerman, 1997, pp. 134-137). A nearly flat trace line across the samples indicates convergence.

4.3.4 Parameter Estimation

With the 10000 samples of parameter values taken from their posterior, the parameter values are estimated by the median value. Likewise, the 95% credibility interval is taken between the 2.5th percentile and the 97.5th percentile. Because values of $E(X_i|\theta)$ were also computed for each of these samples, their posterior expectations and credibility intervals are also reported. In the application described below, the 10000 samples were obtained after 10000 burn-in iterations by running the Gibbs sampler another 100000 iterations and retaining each 10th draw.

4.3.5 Assessing Model Adequacy

Three statistics are considered to assess the fit of the M and IIO constrained LCMs. The first one is the deviance information criterion (DIC; Spiegelhalter, Best, Carlin, & Van der Linde, 2002), which can be used to compare the overall goodness-of-fit of competing models. In addition, two statistics were developed to assess assumptions M and IIO, respectively, at the item level.

Bayesian Deviance Statistic

Under inequality constraints, the number of free parameters of a model (or model complexity) is not easily defined. However, Spiegelhalter, Best, Carlin, and Van der Linde (2002) proposed a Bayesian measure p_D for the effective number of parameters of a model that can easily be obtained with our Gibbs sampler. Let $P(\mathbf{X}|\pi)$ denote the likelihood function, and let $D(\pi) = -2\ln P(\mathbf{X}|\pi)$ denote the deviance. Using the sampled parameter values from the Gibbs sampler, the following two quantities can be computed: $P_E\big(D(\pi)\big)$ which is the posterior expectation of the deviance, and $D\big(P_E(\pi)\big)$ which is the deviance given the posterior expectation of the parameters. The effective number of parameters is expressed as

$$p_D = P_E\big(D(\pi)\big) - D\big(P_E(\pi)\big).$$

Spiegelhalter, Best, Carlin, and Van der Linde (2002) suggested using DIC as a measure for comparing model-data fit across competing models. Similar to Akaike's (1973) information criterion, DIC equals the deviance penalized by the complexity of the estimated model. It is defined as

$$\text{DIC} = P_E\big(D(\pi)\big) + p_D.$$

Lower DIC values indicate a better model-data fit.

Posterior Check for Monotonicity

Now we define the statistic expressing the deviation from M for item i. For two items i and i', let the corresponding *rest score* be defined as

$$Y_{ii'} = \sum_{i'' \neq i,i'} X_{i''},$$

which is the sum score excluding items i and i'. Moreover, let $S(X_i, X_{i'}|y)$ be the covariance between X_i and $X_{i'}$ conditional on their rest score $Y_{ii'} = y$. Holland and Rosenbaum (1986; Rosenbaum, 1984) showed that M implies that these covariances are nonnegative. Let n_y denote the number of observations with rest score $Y_{ii'} = y$. We define the item specific statistic for M as

$$M_i(\mathbf{X}) = \frac{1}{n} \sum_{i' \neq i} \sum_{y} n_y S(X_i, X_{i'}|y), \tag{8}$$

where the weighing factor n_y accounts for small observations yielding less reliable estimates of the covariances.

To obtain a distribution of the statistic in Equation 8 under the null-hypothesis that the constrained LCM holds, new data sets $\mathbf{X}^{(t)}$ are generated using the parameter values of the tth sample. For each of these data sets, statistic $M_i(\mathbf{X}^{(t)})$ is computed, which yields a sampling distribution for the test statistic under the null-hypothesis of the constrained LCM. We define the (one-sided) p-value for assessing for each item whether it fits the constrained LCM as the proportion of times $M_i(\mathbf{X}^{(t)})$ is smaller than the observed $M_i(\mathbf{X})$ (cf. Gelman, Meng, & Stern, 1996; Meng, 1994). A small p-value undermines the credibility of assumption M for that particular item.

Posterior Check for Invariant Item Ordering

The last statistic expresses the deviation from IIO for item i. It is derived in a similar manner as $M_i(\mathbf{X})$. For two items i and i', where $i < i'$, IIO implies that $E(X_{i'} - X_i|y) \geq 0$, for all $Y_{ii'} = y$ (Ligtvoet, Van der Ark, Bergsma, &

Sijtsma, 2009). As a statistic for goodness of fit of item i, we propose

$$IIO_i(\mathbf{X}) = \frac{1}{n} \sum_{i' \neq i} \sum_{y} n_y E(X_{i'} - X_i|y). \tag{9}$$

In Equation 9, we only consider for i' the items adjacent to i. Similar to $M_i(\mathbf{X})$, a posterior sample distribution of $IIO_i(\mathbf{X})$ with the corresponding p-value can be computed by replicating new data sets $\mathbf{X}^{(t)}$. A small p-value undermines the credibility of the IIO assumption for that particular item.

Model Fitting Strategy

We propose using a three-step model fitting strategy for assessing whether assumptions M and IIO hold. The steps involve: 1) determining the number of latent classes $\hat{q}$, 2) determining for which items M holds, and 3) determining for which items IIO holds. This procedure yields a LCM for which M and IIO holds for a certain number of items. We refer to the number of items for which M holds as $\hat{k}^M$ and for which IIO holds as $\hat{k}^{IIO}$.

Step 1 involves estimating LCMs with different numbers of classes, where the model with the lowest DIC is retained for the next step. Step 2 starts by fitting a M restricted $\hat{q}$-class model and comparing its DIC value to the unrestricted $\hat{q}$-class model. In case the constrained LCM fits worse, the posterior check for M is used to determine for which items assumption M is violated. The M constraint is then removed for the item with the largest misfit, and the model is re-estimated with M constraints on the remaining items. This is repeated until the DIC indicates that the LCM constrained by M for the remaining $\hat{k}^M$ items fits at least as well as the unconstrained LCM. Step 3 starts with the estimation of a LCM constrained by both M and IIO, with constraints imposed only on those $\hat{k}^M$ items for which M held. The fit of this LCM is compared to the fit of the final model from step 2. As long as the DIC values indicate that the IIO restricted model fits worse, the IIO constraint is relaxed for the item for which the posterior check indicates the largest misfit due to IIO.

4.4 Application

To illustrate the procedure for testing M and IIO by means of constrained LCMs, we use an application to a set of items from a study on sociocultural developments in The Netherlands (Felling, Peters, & Schreuder, 1987; Heinen, 1996, chap. 2 and 3). These were self-ratings of 1134 subjects to five statements related to women's liberation, each with three ordered score categories.

Because DIC values showed that an unrestricted four-class model did not improve the model-data fit compared to a three-class model, we retained the three-class model for testing M and IIO. The unconstrained LCM with three

Table 4.1: Median Values Released Constrained LCM.

Latent	Class	Conditional Expectation				
Class	Proportion	Item 1	Item 2	Item 3	Item 4	Item 5
1	.16	.23	.17	.55	.69	1.49
2	.34	.72	.73	1.30	1.65	1.90
3	.51	1.61	1.79	1.69	1.93	1.97

latent classes had a DIC equal to 8539.62. The estimates of the conditional expectations revealed that the assumption M held for all items, whereas the item ordering of items 1 and 2 was different at the first latent class than at the remaining two classes, and the item ordering of items 2 and 3 was different at the third latent class.

Then a LCM was fitted constrained by M to all five items. This constrained LCM fitted as well as the unconstrained LCM, with DIC = 8539.56. The posterior check for M also indicated that all items fitted the model. So there is no reason to relax the M assumption for one of the items.

With the LCM constrained by both M and IIO for all five items we obtained DIC = 8544.52, which indicated that the model fitted the data less well than the LCM constrained by only M. The posterior check for IIO indicated that the lack of model-data fit might be due to item 2, for which $IIO_2 = 0.284$ ($p = .001$). All other items had p-values larger than .10. The posterior checks for M remained similar to those of the LCM constrained by only M. Then we released the IIO constraint for item 2, which means that item 1 is now constrained by IIO from above by item 3 and item 3 from below by item 1. This model showed a good model-data fit based on DIC = 8538.93 (lowest value of DIC for all models tested). Non of the posterior checks for M and IIO showed any item misfit. Table 4.1 contains the median values of the samples of the class proportions and conditional expectations under the LCM constrained by M for all five items and IIO for items 1, 3, 4, and 5.

4.5 Discussion

It was shown how the M and IIO assumptions in IRT models for ordinal items can be translated into LCMs with inequality constraints on the class-specific item means. A Gibbs sampling procedure was presented which uses a re-parametrization of the model probabilities to make the implementation of the complex inequality constraints feasible. Our data application illustrated the proposed three-step model fitting strategy for determining the set of items for which M and IIO holds. This procedure makes use of the overall fit statistic DIC, as well as item-specific posterior checks.

Because most IRT models are used with the goal of obtaining an ordering of subjects, it is most logical to assume IIO only in addition to M, which is also what we did in this chapter. However, in theory it is also possible to test IIO without assuming M. This could be done using a LCM with only the IIO constraints. A practical problem one may encounter when estimating such a model with a Gibbs sampler is what is referred to as the label switching problem (e.g., Redner & Walker, 1984; Stephens, 2000). Because the classes are not ordered, their labeling is arbitrary and may thus change during the Gibbs sampling iterations (see Hoijtink, 1998, for an example of this phenomena). Note that this may also occur in an unconstrained LCM. Stephens (1997, 2000) developed an algorithm that can be used to reorder the classes so that their order is same across Gibbs sample iterations (see also Jasra, Holmes, & Stephens, 2005).

Two posterior checks were proposed for assessing which items violate M or IIO when DIC indicates that the specified model does not hold. In the data application with five items and 1134 subjects, the posterior checks seem to work well. However, the sample size required for these checks to achieve a reasonable power remains a question for future research. Moreover, alternative checks may be developed.

The proposed sequential approach for testing the assumptions M and IIO did not exhaustively assess all possible combinations of items for which either M and IIO holds. For example, in our application, all items corresponded well to M, whereas the IIO constraint was removed for item 2. However, because at least two items are involved in a violation of IIO, it might be the case that releasing the IIO constraint for either item 1 or 3 would have resulted in a better fitting model. Future research should show whether or not our model fitting strategy is optimal in selecting the best fitting model.

Also, in our analysis the number of latent classes was fixed to three based on the goodness-of-fit of the unrestricted LCM. An alternative strategy for testing M and IIO using LCMs might be to use a model with a large number of latent classes as the starting point. However, a downside of this alternative strategy is that some of the classes will contain only a few subjects, hence reducing the reliability of the estimates. Another possibility may be to treat the number of latent classes as an additional parameter in the Bayesian estimation procedure. This would yield a Gibbs sampler with in addition to the augmentation and the two sampling steps, a step in which the number of latent classes is updated (e.g., Neal, 1991; Richardson & Green, 1997).

A last extension of the proposed procedures for testing M and IIO we would like to mention is the possibility of imposing constraints that correspond to more common IRT models implying M and IIO. For example, Van Onna (2002) defined a LCM with order restrictions similar to ours but imposed on the cumulative probabilities $P(X_i \geq x_i|\theta)$ instead of the expected values $E(X_i \geq x_i|\theta)$. These restrictions imply both M and IIO to hold (cf. the strong double monotonicity model, Sijtsma & Hemker, 1998). Ligtvoet, Van der Ark, Bergsma and

Sijtsma (2009) suggested models with similar order restrictions for continuation ratios and adjacent category odds (see also Vermunt, 2001).

Chapter 5

On the number of items that can be invariantly ordered in samples

Abstract

Items in a test that have the same ordering according to their difficulty across all ability levels exhibit an invariant item ordering (IIO). Provided IIO holds in the population, the reliable estimation of an item ordering may be problematic in small samples, but also in samples of realistic size. For a given sample size, the gravity of this problem depends on the shape of the item response functions and the distribution of the latent variable. It is shown that, under ideal circumstances, in samples of realistic size no more than approximately six items can be ordered reliably so that IIO can be inferred for these items. Another result is that, for the investigation of IIO in a sample of realistic size, a discretization of the latent variable into more than nine ordered discrete categories leads to a loss of accuracy that is too big to be able to infer IIO.

5.1 Introduction

Items in a test for children's intelligence assessment are often administered in an ordering from easy to difficult. As a starting rule, children skip the items that are considered too easy for their age level. The stopping rule is that the testing procedure terminates when the child fails, say, three consecutive items. Norm tables can be constructed to assess the performance of each child within his/her age category. An example of a test that uses this administration procedure is the Wechsler Intelligence Scale for Children (WISC; Wechsler, 2003).

The administration procedure requires the items to have the same ordering by increasing item difficulty across all age categories, but also within each age category across all levels of the intelligence scale. Such an item ordering is called invariant item ordering (IIO). IIO means that the items in a test have the same ordering according to their difficulty across all ability levels. Other applications for which IIO is relevant are person-fit analysis (Emons, Sijtsma, & Meijer, 2004), differential item functioning (Shealy & Stout, 1993), and the testing of theories about developmental processes (Bouwmeester & Sijtsma, 2007).

Let k denote the number of items in the test, and let X_i be the item score variable of item i ($i = 1, \ldots, k$), with realizations $x_i \in \{0, 1\}$. In item response theory (IRT) models, often a unidimensional latent variable, denoted θ, is assumed to account for the associations between the item scores. Conditional on θ, the item score variables of all k items are assumed to be independent; this is the assumption of local independence (LI). Let the item response function (IRF) be defined as $P(X_i = 1|\theta)$. IRT models for dominance data (Coombs, 1964, p. 23), such as intelligence test data and other maximum-performance data but also personality assessment data and other typical-performance data, assume IRFs to be monotone nondecreasing or strictly increasing. This is the monotonicity assumption. Because IIO is relevant in the context of IRT models for dominance data, we use monotone IRFs in the examples throughout.

Sijtsma and Junker (1996) defined IIO as follows. A test that consists of k items has an IIO if the items can be ordered and numbered accordingly such that

$$P(X_1 = 1|\theta) \leq P(X_2 = 1|\theta) \leq \ldots \leq P(X_k = 1|\theta), \qquad \text{(IIO)}$$

for all θ. The definition of IIO is such that items become easier as the item index increases. In the WISC example, the inequality signs in the definition of IIO would have to be reversed. IIO allows the possibility that for certain values of θ the ordering contains ties, but implies that the IRFs do not intersect. Figure 5.1a shows three IRFs for which IIO holds, including ties. Figure 5.1b shows three IRFs for which IIO does not hold; the solid IRF intersects once with the dotted IRF and twice with the dashed IRF.

The concept of IIO also appears elsewhere in the psychometric literature. Rosenbaum (1987a) defined a latent scale as a set of items for which LI and IIO

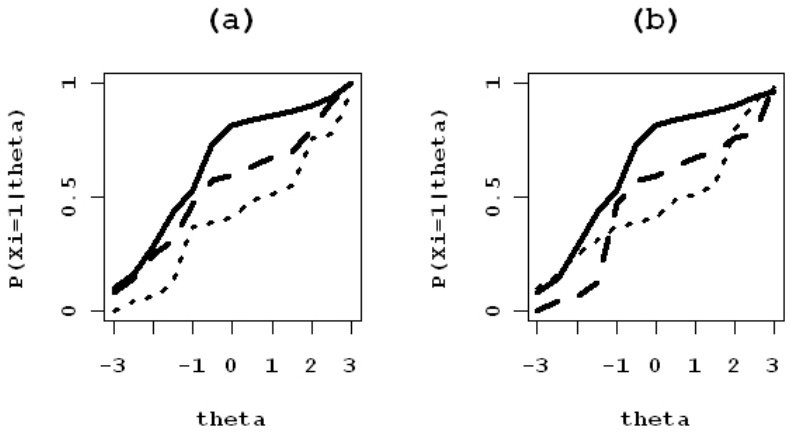

Figure 5.1: Two examples of three IRFs: (a) IIO holds, and (b) IIO is violated.

hold. In another context, Rosenbaum (1987b) defined item i to be uniformly more difficult than item j if $P(X_i = 1|\theta) \leq P(X_j = 1|\theta)$ for all θ. This ordering property is equivalent to characteristic monotonicity in unfolding models (Ellis, 1994, chap. 6). Examples of IRT models that imply an IIO are the double monotonicity model (Mokken, 1971, p. 118), the Rasch model (Rasch, 1960), and the isotonic ordinal probabilistic model (Scheiblechner, 1995). Examples of IRT models that do not imply an IIO are the normal ogive model (Lord, 1980, pp. 30-32) and the 2- and 3-parameter logistic models (Birnbaum, 1968).

Suppose, k items have a strict IIO in the population of interest; that is, IIO holds with strict inequalities. Let b_i denote the difficulty parameter of item i, and let the Rasch (Rasch, 1960; Fischer, 1974, 1995) model define the IRF of item i as a logistic function

$$\text{logit}P(X_i = 1|\theta) = \theta - b_i. \tag{1}$$

The Rasch model implies IIO. An example of two items with a *strict* IIO is obtained by taking for two items i and j in Equation 1 the difficulty parameters $b_i > b_j$. This paper investigates the problem illustrated in Figure 5.2. A stratified sample of size $n = 300$ was drawn from θ intervals $(-\infty, -2], (-2, -1.5], \ldots, (1, 1.5], (1.5, \infty)$ based on $\theta \sim N(0, 1)$. The scores on the two items i and j were simulated for $b_i = 1$ and $b_j = -1$ in Equation 1. For these scores, the confidence intervals were estimated for both items at each increment of θ (here, θ was assumed to be known). Figure 5.2 shows for the two items at each of the 5 θ-values the estimated proportion (black diamonds) and their 95% confidence intervals. Note that, for the estimation of the probabilities, we did not assume the Rasch model.

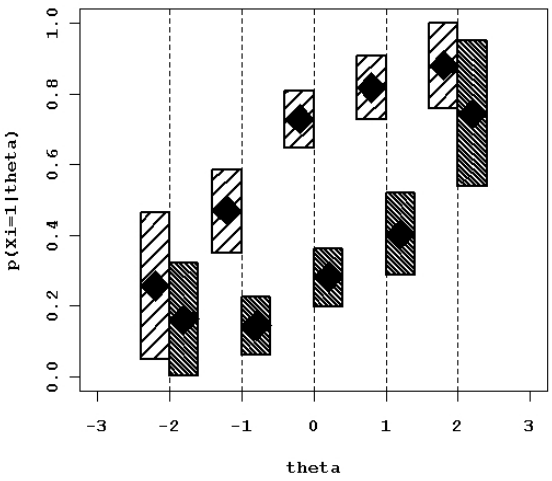

Figure 5.2: The 95% confidence intervals based on a sample of size 300.

The example in Figure 5.2 shows that at $\theta = -2$ and $\theta = 2$ IIO could not be reliably inferred as the confidence intervals overlap considerably. Hence, a sample size of 300 does not allow estimates of the two IRFs that are reliable enough to infer that the IRFs do not intersect. Further, for a realistic number of items (more than two) this is likely to be a problem even if the sample size is much larger than $n = 300$. The gravity of the problem depends on the interplay of three factors: the constellation of the k IRFs, the distribution of θ, and the sample size n.

This study investigates the number of items of which the IRFs can be distinguished accurately in a sample, so as to infer correctly that IIO holds in the population of interest. The point of departure is strict inequality in the definition of IIO, which corresponds to an item ordering where the items are distinguishable. The first research question is: Given a strict IIO and a sample size n, what is the maximum number of IRFs that can be distinguished significantly? The second research question is: Given a strict IIO, what is the minimally required sample size needed to significantly distinguish k IRFs? The two research questions were investigated separately, and the results are discussed subsequently in Study 1 and Study 2.

5.2 Study 1

In Study 1, the first research question was investigated: Given a strict IIO and a sample size n, what is the maximum number of IRFs that can be distinguished

significantly? The number of IRFs was investigated for one arbitrary value of θ. Notice that for one value θ_0, the IRFs reduce to response probabilities $P(X_i = 1|\theta_0)$. This allows the investigation of the effect of sample size at θ_0 (denoted n_0), without having to consider the shape of the IRFs or the distribution of θ. The results of Study 1 are used in Study 2 to define a set of IRFs.

5.2.1 Method

For θ_0, two response probabilities for adjacent items, $P(X_i = 1|\theta_0)$ and $P(X_{i+1} = 1|\theta_0)$, can be significantly distinguished if the null hypothesis that

$$\mathcal{H}_0 : \; P(X_i = 1|\theta_0) = P(X_{i+1} = 1|\theta_0) \tag{2}$$

is rejected in favor of the alternative hypothesis that

$$\mathcal{H}_A : \; P(X_i = 1|\theta_0) < P(X_{i+1} = 1|\theta_0). \tag{3}$$

If a strict IIO holds, the rejection of $\mathcal{H}_0$ implies that two *non*adjacent response probabilities, such as $P(X_i = 1|\theta_0)$ and $P(X_{i+2} = 1|\theta_0)$, can also be significantly distinguished. Thus, for a strict IIO of k items to hold it is sufficient to reject $\mathcal{H}_0$ for all adjacent pairs. The maximum number of response probabilities that can be distinguished depends on the sample size n_0, the power $1 - \beta$, and the nominal Type I error α.

Let k^* denote the maximum number of response probabilities that can be distinguished significantly. The procedure for determining k^* at θ_0 has three steps which are discussed below and illustrated in Figure 5.3.

1. In deriving k^*, we use normal approximations to the difference between $\hat{P}(X_i = 1|\theta_0)$ and $\hat{P}(X_{i+1} = 1|\theta_0)$ (Appendix A) which are not valid if values of $P(X_i = 1|\theta_0)$ and $P(X_{i+1} = 1|\theta_0)$ are close to zero or unity. To prevent the response probabilities from being close to zero and unity, a lower limit l and an upper limit u are chosen such that $0 < l < u < 1$. In this study,

$$l = \frac{5}{n_0}, \text{ and } u = \frac{n_0 - 5}{n_t}.$$

   The first response probability $\hat{P}(X_1 = 1|\theta_0)$ is set equal to l (Figure 5.3a). The nominal Type I error is set to α.

2. The response probabilities of the items $\hat{P}(X_2 = 1|\theta_0), \ldots, \hat{P}(X_{k^*+1} = 1|\theta_0)$ are computed consecutively. Let d_i denote the smallest difference between $\hat{P}(X_i = 1|\theta_0)$ and $\hat{P}(X_{i+1} = 1|\theta_0)$ for which $\mathcal{H}_0$ is rejected. Once $\hat{P}(X_i = 1|\theta_0)$ $(i = 1, \ldots, k^*)$ has been estimated, $\hat{P}(X_{i+1} = 1|\theta_0)$ is computed as $\hat{P}(X_{i+1} = 1|\theta_0) = \hat{P}(X_i = 1|\theta_0) + d_i$ (Figure 5.3b); d_i

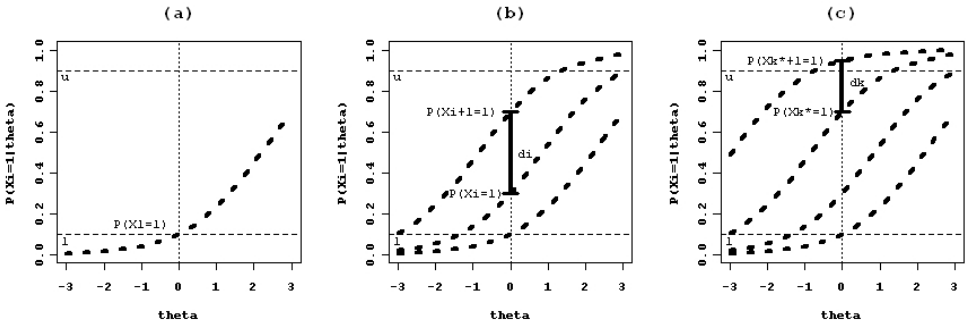

Figure 5.3: Illustration of the procedure for determining the maximum number of response probabilities at θ_0.

is derived in Appendix A. The procedure stops when $\hat{P}(X_{k^*+1} = 1|\theta_0)$ exceeds the upper limit u (Figure 5.3c).

3. In Step 2, k^* statistical tests were conducted. The nominal Type I error is adjusted for multiple testing (Bonferroni correction), such that the adjusted nominal Type I error equals α/k^*. Step 2 is repeated with the adjusted nominal Type I error. An adjustment of the nominal Type I error rate affects d_i and, therefore, it may also affect k^*. Hence, the value of k^* that was obtained the first time Step 2 was run, now denoted $k^{*(1)}$, need not equal k^* that was obtained the second time Step 2 was run, which is denoted $k^{*(2)}$ and, as a result, the nominal Type I error is adjusted to $\alpha/k^{*(2)}$. Step 2 is repeated until after the vth run $k^{*(v)} = k^{*(v-1)}$.

The procedure ensures that the obtained k^* response probabilities are located just far enough apart, such that, if for any two items we were to test $\mathcal{H}_0$ (Equation 2) at θ_0 (given n_0, $1 - \beta$, and α) the null hypothesis would be rejected.

In this study, the dependent variable was the maximum number of response probabilities that can be distinguished at θ_0, and the independent variables were the sample size at n_0 and the power of the statistical test. Sample size n_0 had five levels (i.e., 100, 200, 400, 1000, and 2000), and the power $1 - \beta$ had three levels (i.e., .70, .80, and .90). The nominal Type I error was fixed at $\alpha = .05$.

5.2.2 Results

In general, the number of response probabilities that can be distinguished at θ_0 was small. Table 5.1 shows that a larger sample size produced a larger k^*. A higher power produced a smaller k^*. The values of $\hat{P}(X_i = 1|\theta_0)$ that correspond to the entries of Table 5.1 are given in Appendix B. These values

Table 5.1: Maximum Number of Response Probabilities that Can be Distinguished Reliably at θ_0, for Overall $\alpha = .05$, and Varying Sample Size and Power.

Sample	Power		
Size	0.70	0.80	0.90
100	6	6	5
200	9	8	7
400	13	12	10
1000	20	18	16
2000	28	26	23

show that d_i is smaller for response probabilities close to zero or unity than for response probabilities close to .5.

5.2.3 Discussion

The maximum number of response probabilities that can be distinguished significantly at one value of θ is small. In practice, it is not expected that many subjects have the same value or approximately the same value of θ. So, what was considered here to be a small sample (i.e., $n_0 = 100, 200$) will be difficult to collect in practice, and may actually be a small part of a very large sample if we were to consider a distribution of θ instead of a single value.

The Bonferroni correction was used to correct for the inflation of the nominal Type I error as the number of statistical tests increased. However, the nominal Type II error (β) was not adjusted for the number of tests. This would have resulted in a higher power, and consequently, in fewer response probabilities that can be distinguished significantly.

5.3 Study 2

In Study 1, we determined the maximum number k^* of response probabilities that can be distinguished significantly for a given sample size and a given power. In Study 2, the second research question was investigated: Given a strict IIO, what is the minimally required sample size needed to significantly distinguish k IRFs? The second study takes the shape of the IRFs into consideration.

5.3.1 Method

The Rasch model Equation 1 implies IIO, and was used to investigate the minimally required sample size needed to significantly distinguish k IRFs. The choice of the Rasch model reduced computational effort. Other unidimensional

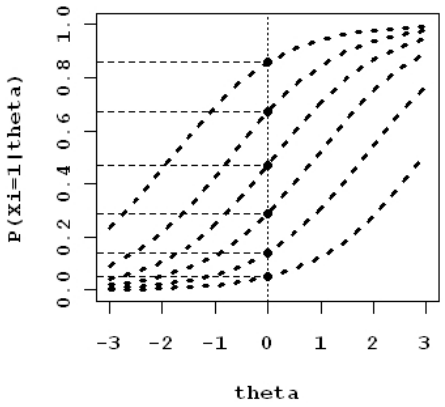

Figure 5.4: Illustration of six IRFs under the Rasch model generated using response probabilities found in Study 1 (b_i parameters were -1.80, -.72, .11, .90, 1.79, and 2.94).

IRT models implying IIO are expected to yield similar results. Given the purpose of this study, it is convenient to rewrite Equation 1. Let a_i denote the intercept of item i at θ_0, that is, $a_i = (1 + e^{b_i})^{-1}$. The Rasch model can be written in terms of the item intercepts (instead of the item difficulties). This version of the model can be obtained from Equation 1 by taking $b_i = \ln(\frac{1-a_i}{a_i})$, which results in

$$P(X_i = 1|\theta) = \left(1 + \frac{1 - a_i}{a_i e^{\theta}}\right)^{-1}. \tag{4}$$

The point of departure in Study 2 was the six response probabilities obtained in Study 1, for a sample size $n_0 = 100$ and a power $1-\beta = .70$ (Appendix B). It was assumed that a test contained six Rasch items each with an intercept equal to one of the six response probabilities; that is $a_1 = .050$, $a_2 = .143$, $a_3 = .289$, $a_4 = .473$, $a_5 = .673$, and $a_6 = .858$. The IRFs of the six Rasch items were computed using Equation 4. Figure 5.4 shows the resulting six IRFs. The minimally required sample size to significantly distinguish k adjacent Rasch IRFs at an arbitrary θ value is derived in Appendix C.

In this study, the independent variable was latent variable θ, which had nine values (i.e., $-2, -1.5, -1, \ldots, 1.5, 2$). The dependent variable was the sample size required at θ to significantly distinguish the IRFs of the six Rasch items with power $1 - \beta = .70$ and nominal Type I error $\alpha = .05$.

Table 5.2: Required Sample Size to Distinguish Six Adjacent Rasch IRFs, for $\alpha = .05$ and $1 - \beta = .70$.

Adjacent	Value of θ								
items	-2.0	-1.5	-1.0	-0.5	0.0	0.5	1.0	1.5	2.0
1,2	585	364	230	149	100	72	55	47	45
2,3	435	281	188	132	101	84	79	85	102
3,4	274	189	139	112	100	102	117	149	207
4,5	139	109	94	91	100	124	167	242	367
5,6	54	53	59	74	101	146	223	349	558
Maximum	585	364	230	149	101	146	223	349	558

* Total of the maximum values in each column: $n = 2705$

5.3.2 Results

The minimally required sample sizes for adjacent items are given in Table 5.2. Table 5.2 shows that, for example, to distinguish items 1 and 2 for the given nominal Type I error and power, the largest sample size is required at $\theta = -2$ ($n_{-2} = 585$) and the smallest at $\theta = 2$ ($n_2 = 45$). To distinguish items 5 and 6, the largest sample size is required at $\theta = 2$ ($n_2 = 558$) and the smallest at $\theta = -1.5$ ($n_{-1.5} = 53$). Table 5.2 also shows the sample size required to distinguish all six IRFs at each of the nine values of θ; this is simply the maximum sample size in each column. Under the Rasch model with an ideal constellation of the intercepts, and given an ideal distribution of the sample, a total sample of at least 2705 subjects is required to distinguish six IRFs across nine values of θ (nominal Type I error equal to .05 and power equal to .70). Figure 5.5a-e shows the distribution of subsamples required to significantly distinguish the pairs of IRFs across all nine values of θ. Figure 5.5f shows the U-shape distribution required to distinguish between all item pairs.

5.3.3 Discussion

Study 1 showed that six response probabilities could be significantly distinguished for a sample size of 100, and nominal Type I error equal to .05 and power equal to .70. Study 2 showed that for nine values of θ, a total sample size of 2705 is required to distinguish six Rasch items that correspond to the response probabilities found in Study 1. Because IRFs were closer in the tails of the distribution of θ, larger subsamples were required there to distinguish the IRFs.

Study 2 revealed that for investigating an IIO most values of θ have to be sampled from the extremes of the θ distribution, whereas in real-data analysis this is exactly where the fewest values of θ are located, given that real ability

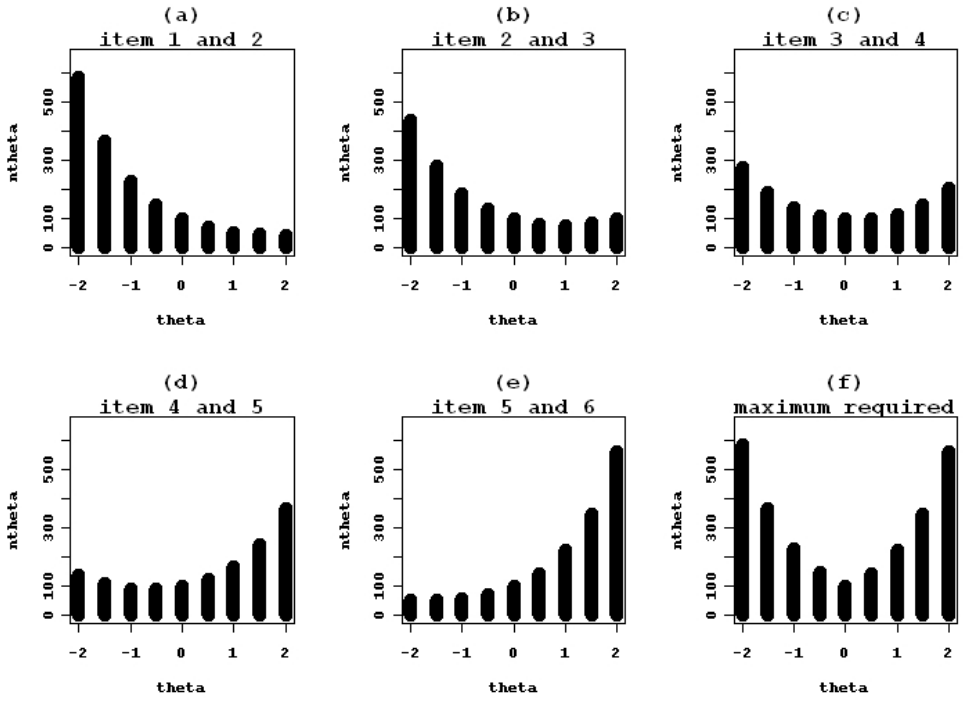

Figure 5.5: Minimally required subsample sizes to distinguish six adjacent Rasch IRFs, for $\alpha = .05$ and $1 - \beta = .70$.

distributions tend to be bell-shaped. This partly explains why samples needed for investigating IIO need to be so large. In practice, when researchers do not know the subjects' values of θ, they have no way to stratify the sampling procedure and, as a result, they sample primarily from the middle of the distribution, simple because the majority of the subjects are located there. However, this drives up sample size in the wrong regions whereas sampling from the tails would be really effective.

Finally, for the investigation of an IIO in a sample of realistic size, a discretization of the latent variable into more than nine ordered categories leads to a loss of accuracy that is too big to infer IIO.

5.4 General discussion

Study 1 showed that, given a strict IIO and a realistic sample size of respondents with the same θ value, and given nominal Type I error α and power $1 - \beta$, the maximum number of IRFs that can be distinguished significantly is at most six. Study 2 showed that, given a strict IIO, the minimally required

sample size based on ideal stratified sampling that is needed to significantly distinguish six IRFs across nine values of θ, is equal to 2705. Under less than ideal circumstances, larger sample sizes are needed to significantly distinguish the same number of items. These circumstances include the unavailability of real values of θ. Rather, values of θ are estimated from the data. Using estimated values of θ has the effect of introducing more uncertainty and requiring even larger samples to reduce this uncertainty to an acceptable level.

The two studies were concerned with situations in which IIO held. In real-data analysis, the researcher often does not have prior knowledge about the ordering of items let alone whether IIO holds for the items. Clearly, IIO is a demanding property of any item set and this is true a fortiori as the number of items grows larger. It may even be reasonable to assume that most item sets do not exhibit IIO in a particular population, so that in real tests applied to real populations IIO is a relatively rare property, that one may only hope to approximate to a degree that serious mistakes in testing subjects are avoided.

What our work also has shown, is that the establishment of IIO (or the lack thereof) requires a large sample of observations at several levels of the latent variable. Given the size of the sample and the typical requirements with respect to the stratification of the sample, it seems safe to say that in practical data analysis samples often will be too small to grant statistical tests enough power for establishing strict IIO. This lack of power implies a large Type II error rate, that is, a large probability of incorrectly assuming that the model (i.e., IIO) is the true model that underlay the generation of the data.

Study 2 further shows that (given IIO and the availability of real values of θ) a sample size of at least 2705 observations is needed to reliably conclude that at most six IRFs are nonintersecting but that drawing the required U-shaped sample distribution will not be easy. The general conclusion from Study 1 and Study 2 is that a coarser, less demanding approach to the study of the response functions is needed for the practical investigation of IIO.

Two alternative approaches that facilitate the investigation of an IIO are the use of clusters of items instead of individual items, and an ordinal latent class approach. First, instead of considering individual IRFs, item clusters may be formed that consist of adjacent IRFs that are more or less similar. Then, an invariant ordering of a limited number of item clusters could be established even if the number of items is large (e.g., Van der Ark & Van Diem, 2003; Verweij, Sijtsma, & Koops, 1999; Wainer, Bradlow, & Wang, 2007; Wainer & Kiely, 1987). Second, instead of assuming θ to be continuous or having more than nine discrete θ levels, a coarser division of the latent variable into ordinal latent classes allows more powerful tests of the ordering of the IRFs. Both approaches may cause more bias (i.e., a more distorted view of the IRFs) but gain precision and keep the sample size realistic.

Future research with respect to an IIO may focus on the requirements of IRFs to be considered similar, how to distinguish different clusters of items, and

the optimal number of latent classes. The theoretical and practical implications of a coarser IRT approach are well worth exploring in view of the practical importance of being able to investigate and establish an invariant orderings of items.

Appendix

A Derivation of d_i

For notational convenience, let $P_i = P(X_i = 1|\theta_0)$, and let $D = \hat{P}_{i+1} - \hat{P}_i$. If $\mathcal{H}_0$ (Equation 2) is true, then for a large sample size n_0 and a value of P_i that is not close to zero or unity, D is approximately normally distributed with mean $E_0(D) = 0$ (subscript 0 refers to the null hypothesis) and standard deviation

$$\begin{aligned} \mathrm{sd}_0(D) &= \sqrt{\frac{P_i(1-P_i) + P_{i+1}(1-P_{i+1})}{n_0}} \\ &= \sqrt{\frac{2P_i(1-P_i)}{n_0}}. \end{aligned} \tag{A1}$$

We used restrictions $P_i > l = n_0/5$ and $P_{i+1} < u = (n_0 - 5)/5$ (e.g., Hoel, 1984, p. 145) to ensure that the response proportions were not close to zero or unity.

For the alternative hypothesis (Equation 3), we consider the simple hypothesis

$$\mathcal{H}_A : \; P_{i+1} - P_i = d_i > 0. \tag{A2}$$

For a large sample size n_0 and values of P_i and P_{i+1} that are not close to zero or unity, D is approximately normally distributed with mean $E_A(D) = d_i$ (subscript A refers to the alternative hypothesis) and standard deviation

$$\begin{aligned} \mathrm{sd}_A(D) &= \sqrt{\frac{P_i(1-P_i) + P_{i+1}(1-P_{i+1})}{n_0}} \\ &= \sqrt{\frac{2P_i(1-P_i) + d_i[1-2P_i-d_i]}{n_0}}. \end{aligned} \tag{A3}$$

Figure 5.6 shows the distributions of D under $\mathcal{H}_0$ (upper panel) and under $\mathcal{H}_A$ (lower panel), and critical value d_z for which $P(D > d_z) = \alpha$. In the upper panel of Figure 5.6 the area under the null distribution that lies right of d_z corresponds to the nominal Type I error α, and the area of the alternative distribution (lower panel) that lies right of d_z corresponds to the power $1 - \beta$. Figure 5.6 shows that d_i consists of the sum of two parts: d_0 and d_A; hence

$$d_i = d_0 + d_A. \tag{A4}$$

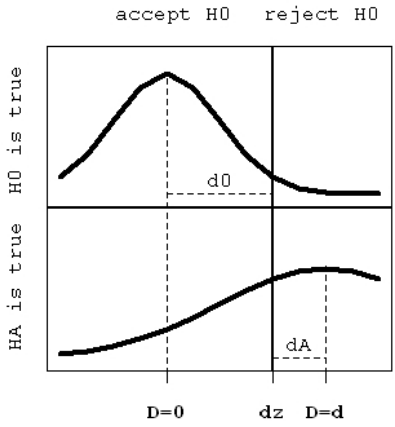

Figure 5.6: Distribution of D under $\mathcal{H}_0$ (upper panel) and under $\mathcal{H}_A$ (lower panel).

Let $z_{1-\alpha}$ be the critical value associated with nominal Type I error α under a standard normal distribution and let z_β be the critical value associated with the nominal Type II error β under a standard normal distribution (note that in our case $z_\beta < 0$); then

$$d_0 = d_{1-\alpha} = z_{1-\alpha} \times \mathrm{sd}_0(D), \tag{A5}$$

and

$$d_A = d_i - d_{1-\alpha} = -z_\beta \times \mathrm{sd}_A(D). \tag{A6}$$

Substituting Equation A5 and Equation A6 into Equation A4 yields

$$d_i = z_{1-\alpha} \times \mathrm{sd}_0(D) - z_\beta \times \mathrm{sd}_A(D), \tag{A7}$$

and substituting Equation A3 into Equation A7 yields

$$d_i = z_{1-\alpha} \times \mathrm{sd}_0(D) - z_\beta \times \sqrt{\frac{2P_i(1-P_i) + d_i(1-2P_i-d_i)}{n_0}}. \tag{A8}$$

Solving d_i in Equation A8 means solving a quadratic equation with solutions

$$d_i = \frac{-B \pm \sqrt{B^2 - AC}}{2A}, \tag{A9}$$

with

$$\begin{aligned} A &= n_0 + z_\beta^2, \\ B &= -2n_0 z_{1-\alpha} \mathrm{sd}_0(D) - z_\beta^2 + 2z_\beta^2 P_i, \text{ and} \\ C &= z_{1-\alpha}^2 n_0 \mathrm{sd}_0^2(D) - 2z_\beta^2 P_i(1-P_i). \end{aligned}$$

Only the solution

$$d_i = \frac{-B + \sqrt{B^2 - AC}}{2A}$$

in Equation A9 yields nonnegative values for d_i.

B Location of the IRFs at θ_0 for all k^* items corresponding to the results of Table 5.1

Sample	Power								
Size	0.70			0.80			0.90		
100	.050	.289	.673	.050	.327	.758	.050	.375	.848
	.143	.473	.858	.157	.539	.940	.175	.616	
200	.025	.270	.704	.025	.306	.780	.025	.360	.876
	.075	.406	.839	.082	.460	.911	.092	.538	
	.158	.554	.947	.177	.623		.206	.718	
400	.013	.323	.851	.013	.371	.923	.013	.437	
	.039	.428	.929	.043	.490	.984	.048	.571	
	.084	.539	.984	.095	.611		.111	.703	
	.148	.650		.169	.729		.199	.823	
	.228	.756		.263	.835		.310	.921	
1000	.005	.259	.754	.005	.298	.830	.005	.357	.918
	.016	.324	.818	.018	.372	.889	.020	.443	.965
	.036	.394	.874	.040	.450	.938	.047	.532	
	.064	.466	.923	.073	.530	.975	.087	.621	
	.101	.540	.961	.116	.610		.139	.706	
	.146	.614	.989	.169	.688		.203	.787	
	.199	.686		.230	.762		.276	.858	
2000	.003	.267	.782	.003	.309	.857	.003	.369	.938
	.008	.315	.827	.009	.363	.897	.010	.432	.968
	.019	.365	.868	.021	.420	.932	.024	.497	.990
	.034	.418	.904	.038	.478	.961	.046	.562	
	.053	.471	.936	.062	.537	.983	.074	.626	
	.078	.525	.962	.091	.595	.997	.109	.689	
	.108	.579	.982	.125	.653		.151	.748	
	.142	.633	.996	.165	.709		.198	.804	
	.180	.685		.209	.762		.251	.855	
	.222	.735		.257	.811		.308	.900	

C Required sample size n_θ to significantly distinguish adjacent Rasch IRFs

For notational convenience, let $P_i = P(X_i = 1|\theta)$. From Equation 4, the conditional variance of item i in terms of intercepts can be expressed for a given value of θ as

$$P(X_i = 1|\theta)\big(1 - P(X_i = 1|\theta)\big) = \frac{e^\theta a_i(1 - a_i)}{\big(a_i(e^\theta - 1) + 1\big)^2}. \tag{C1}$$

The difference between two adjacent Rasch IRFs can be expressed for a arbitrary value of θ. Let d_i denote this difference. From Equation 4 we obtain

$$\begin{aligned} d_i &= P_{i+1} - P_i \\ &= \left(1 + \frac{1 - a_{i+1}}{e^\theta a_{i+1}}\right)^{-1} - \left(1 + \frac{1 - a_i}{e^\theta a_i}\right)^{-1} \\ &= \frac{e^\theta a_{i+1}}{a_{i+1}(e^\theta - 1) + 1} - \frac{e^\theta a_i}{a_i(e^\theta - 1) + 1} \\ &= \frac{e^\theta(a_{i+1} - a_i)}{\big(a_{i+1}(e^\theta - 1) + 1\big)\big(a_i(e^\theta - 1) + 1\big)}. \end{aligned} \tag{C2}$$

In Appendix A the standard deviation of the sample statistic D under $\mathcal{H}_0$ is given in Equation A1. Substituting Equation A1 in Equation A8 yields

$$\begin{aligned} d_i &= \frac{1}{\sqrt{n_\theta}}\left(z_{1-\alpha}\sqrt{2P_i\big(1 - P_i\big)} - \right. \\ &\quad \left. z_\beta\sqrt{2P_i\big(1 - P_i\big) + d_i\big(1 - 2P_i - d_i\big)}\right) \quad \Leftrightarrow \\ n_\theta &= d_i^{-2}\left(z_{1-\alpha}\sqrt{2P_i\big(1 - P_i\big)} - \right. \\ &\quad \left. z_\beta\sqrt{2P_i\big(1 - P_i\big) + d_i\big(1 - 2P_i - d_i\big)}\right)^2 \end{aligned} \tag{C3}$$

Now, by substituting Equation 4, Equation C1, and Equation C2 in Equation C3, the minimal required sample size n_θ is obtained that is needed to significantly distinguish adjacent Rasch IRFs at an arbitrary value of θ for a nominal Type I error α, power $1 - \beta$, and item intercepts a_i and a_{i+1}.

Chapter 6

Polytomous latent scales for the investigation of the ordering of items*

Abstract

Latent scales are defined within the framework of nonparametric item response theory for polytomously scored items. Latent scales imply an invariant item ordering, without imposing parametric restrictions on the shape of the item response functions. A hierarchical relationship between three latent scales is derived. The observable properties of manifest invariant item ordering, manifest-scale cumulative probability model, and increasingness in transposition are derived. A real-data† example illustrates the investigation of latent scales by means of these manifest properties.

*This Chapter has been submitted for publication.

†The authors would like to thank Bas Hemker for providing data.

6.1 Introduction

Nonparametric item response theory (IRT) imposes only restrictions on the model structure that are necessary for obtaining useful measurement properties. For example, Ünlü (2007, 2008; Grayson, 1988; Hemker, Sijtsma, Molenaar, & Junker, 1997; Huyhn, 1994) studied the assumptions that are necessary for ascertaining the stochastic ordering of subjects on the latent variable scale by means of the observed total score on the items comprising a test. Another example, in particular relevant to this study, is an ordering of items by difficulty or unpopularity that is the same for all subject measurement values. Such an item ordering is an *invariant item ordering* (IIO; Sijtsma & Junker, 1996; Sijtsma & Hemker, 1998). The purposes of this study are to present and discuss new models for IIO also known as latent scales (Rosenbaum, 1987a), and methods to investigate the fit of these latent scales to test data.

IIO is necessary or desirable in several test applications (Sijtsma & Molenaar, 2002, pp. 92-96). Areas of application in the cognitive domain are intelligence testing, in which the items are often administered in the order from easy to difficult, and starting and stopping rules are based on the alleged (but rarely empirically investigated and supported) invariant ordering of the items (e.g., Wechsler, 2003). Another area is the testing of developmental sequences assumed to be the same for all subjects and represented by different items that require processes and skills typical for a particular phase but not for others (e.g., Jansen & Van der Maas, 1997). Finally, in person-fit analysis item-score patterns of subjects may be assessed to be aberrant because they deviate from an expected common item ordering (e.g., Meijer & Sijtsma, 2001).

In the typical-behavior domain, researchers often assume that items have an invariant cumulative structure, reflecting a hierarchy of psychological or physical symptoms (e.g., Van Schuur, 2003; Watson, Deary, & Shipley, 2008). For example, a rating scale statement like "*I do not talk a lot in the company of other people*" seems to invite higher ratings than "*I prefer not to see people and do things on my own*", because the former statement seems to refer to a

less intense symptom of introversion. Questionnaires for cognitive Alzheimer symptoms assume a fixed order in which these symptoms manifest themselves (Ligtvoet, Van der Ark, & Sijtsma, 2008).

For dichotomously scored items, IRT models implying IIO are the nonparametric Mokken (1971) double monotonicity model and its special case, the Rasch (1960) model. IIO is a property of these models, which in addition have other assumptions. Glas and Verhelst (1995) discussed methods for investigating goodness of fit of the Rasch model, and Sijtsma and Junker (1996) for investigating IIO in a nonparametric IRT context. For polytomously scored items, IIO is inconsistent with the common IRT models, and only a few, restrictive IRT models imply IIO (Sijtsma & Hemker, 1998). Some methods are available for investigating IIO in polytomous data (Ligtvoet, Van der Ark, Te Marvelde, & Sijtsma, in press). One of these methods is at the basis of a method discussed later.

In the next sections, we define IIO, discuss three classes of IRT models for polytomously scored items, explain why most polytomous IRT models do not imply IIO, and implement a sufficient condition for the IIO property in the three different classes of polytomous IRT models. We prove that the three classes of models are hierarchically related, and that all three imply the IIO property. We derive observable consequences, propose different methods for investigating these consequences in real data, and illustrate the methods in two real-data examples.

6.2 Invariant Item Ordering and Polytomous IRT Models

Let a test consist of k items, indexed by $i = 1, \ldots, k$. Let random variable X_i denote the item score; X_i has ordered realization x ($x = 0, \ldots, m$). These scores may reflect the degree to which a subjects has solved a cognitive item correctly or endorsed a typical-behavior statement presented in a rating scale item. The unidimensional latent variable is denoted by θ, and represents the cognitive

ability or the personality trait of interest. Finally, let $E(X_i|\theta)$ be the conditional expectation of the item score X_i, also known as the item response function (IRF; Chang & Mazzeo, 1994). For dichotomously scored items with $x = 0, 1$, we have that $E(X_i|\theta) = P(X_i = 1|\theta)$, which is the conditional probability of answering correctly to item i.

For polytomously scored items, Sijtsma and Hemker (1998) defined IIO as follows. A set of k items with $m + 1$ ordered answer categories per item have an invariant item ordering (IIO) if the items can be ordered and numbered accordingly such that

$$E(X_1|\theta) \leq E(X_2|\theta) \leq \ldots \leq E(X_k|\theta), \tag{IIO}$$

for all θ. It may be noted that IIO allows for ties, so that for some values of θ the item ordering is partial.

6.2.1 Three Classes of Polytomous IRT Models

The three classes of polytomous IRT models are the *cumulative probability models*, *continuation ratio models*, and *adjacent category models* (Agresti, 1990; Hemker, Van der Ark, & Sijtsma, 2001; Mellenbergh, 1995; Molenaar, 1983). The classes each assume unidimensionality, and local independence; that is, for a k-dimensional vector of item scores $\mathbf{X} = \mathbf{x}$,

$$P(\mathbf{X} = \mathbf{x}|\theta) = \prod_{i=1}^{k} P(X_i = x|\theta). \tag{1}$$

An item having $m + 1$ ordered answer categories has m *item steps*, which have to be passed in going from category 0 to category m (Molenaar, 1983). The probability of passing the item step conditional on θ is the *item step response function* (ISRF). The three classes of IRT models differ in their definition of the ISRF, and models within classes place different restrictions on their class-specific ISRF.

Cumulative probability models (CPMs) define ISRFs as

$$\begin{aligned} C_{x_i}(\theta) &= P(X_i \geq x|\theta) \\ &= \sum_{u=x}^{m} P(X_i = u|\theta), \end{aligned} \tag{2}$$

for $x = 1, \ldots, m$, and $C_{x_i}(\theta) = 1$ for $x < 1$, and $C_{x_i}(\theta) = 0$ for $x > m$. This ISRF definition implies that the ISRFs of the same item cannot intersect (Mellenbergh, 1995). Examples of CPMs are the homogeneous case of the graded response model (Samejima, 1969, 1997), and the nonparametric graded response model (Hemker et al., 1997; also, see Molenaar, 1997). These models assume that the ISRF defined by $C_{x_i}(\theta)$ (Equation 2) increases monotonically (i.e., the monotonicity assumption). Van Engelenburg (1997, chap. 2, 3) argued that CPMs are particulary suited for modeling item scores that result from a global assessment task as with rating scales.

Continuation-ratio models (CRMs) define ISRFs as

$$\begin{aligned} M_{x_i}(\theta) &= P(X_i \geq x|X_i \geq x-1;\theta) \\ &= \frac{\sum_{u=x}^{m} P(X_i = u|\theta)}{\sum_{v=x-1}^{m} P(X_i = v|\theta)}, \end{aligned} \tag{3}$$

for $x = 1, \ldots, m$, and $M_{x_i}(\theta) = 1$ for $x < 1$, and $M_{x_i}(\theta) = 0$ for $x > m$. Examples of CRMs are the sequential Rasch model (Tutz, 1990), and the nonparametric sequential model (Hemker et al., 2001). These models assume monotonicity for $M_{x_i}(\theta)$ (Equation 3). Items typically suited for CRM analysis consist of m subtasks that have to be executed in a fixed order such that failing a subtask implies failing the next subtasks, and the item score reflects that the first x subtasks have been succeeded (Hemker et al., 2001).

Adjacent category models (ACMs) define ISRFs as

$$\begin{aligned} A_{x_i}(\theta) &= P(X_i = x | X_i = x \vee X_i = x - 1; \theta) \\ &= \frac{P(X_i = x|\theta)}{P(X_i = x - 1|\theta) + P(X_i = x|\theta)}, \end{aligned} \tag{4}$$

for $x = 1, \ldots, m$, and $A_{x_i}(\theta) = 1$ for $x < 1$, and $A_{x_i}(\theta) = 0$ for $x > m$. Examples of ACMs include the rating scale model (Andrich, 1978), the partial credit model (Masters, 1982; Masters & Wright, 1997), the generalized partial credit model (Muraki, 1992), and the nonparametric partial credit model (Hemker et al., 1997). These models assume monotonicity for $A_{x_i}(\theta)$ (Equation 4). Van Engelenburg (1997, p. 38) noticed that the structure of ACMs does not correspond well with a particular task structure but he suggested that ACMs are best suited for analyzing item scores that result from tasks that consist of x subtasks, which may be solved in an arbitrary order. An item score of x means that any x subtasks were solved correctly.

Van der Ark, Hemker and Sijtsma (2002) showed that the mathematically most general representatives of each of the three classes, which are the nonparametric graded response model (CPM class), the nonparametric sequential model (CRM class), and the nonparametric partial credit model (ACM class) have a hierarchical relationship; that is, using obvious acronyms,

$$\text{np-PCM} \;\Rightarrow\; \text{np-SM} \;\Rightarrow\; \text{np-GRM}.$$

Thus, $A_{x_i}(\theta)$ (ACM class) provides the strongest form of monotonicity, and $C_{x_i}(\theta)$ (CPM class) the weakest. For dichotomously scored items, the three classes coincide, such that $C_{x_i}(\theta) = M_{x_i}(\theta) = A_{x_i}(\theta) = P(X_i = 1|\theta)$.

6.2.2 Relating IIO and Polytomous IRT Models

Sijtsma and Hemker (1998) showed that well-known polytomous IRT models such as the partial credit model (Masters, 1982) and the graded response model (Samejima, 1997) do not imply IIO. The few parametric models that imply

IIO are among the most restrictive IRT models known in the IRT literature, such as the rating scale model (Andrich, 1978) and a restricted version of Muraki's (1990) rating scale version of the graded response model (Sijtsma & Hemker, 1998; Van der Ark, 2001). These models are highly restrictive and typically fail to fit data well. Hence, there appears to be a mismatch between polytomous IRT models and the IIO property. The mismatch is due to the focus of polytomous IRT models on sets of more-detailed ISRFs rather than the aggregate IRF, $E(X_i|\theta)$, on which the definition of IIO is based. For example, the ISRFs of the class of CPMs (Equation 2) are related to the IRFs by

$$E(X_i|\theta) = \sum_{x=1}^{m} C_{x_i}(\theta). \tag{5}$$

We use Samejima's (1997) homogeneous case of the graded response model as an example of the mismatch between ISRF and IRF. This model defines the ISRF as a monotone increasing logistic function with slope parameter $\alpha_i > 0$ and score-category location parameter β_{x_i} $(x_i = 1, \ldots, m)$, as

$$C_{x_i}(\theta) = \frac{e^{\alpha_i(\theta-\beta_{x_i})}}{1 + e^{\alpha_i(\theta-\beta_{x_i})}}. \tag{6}$$

Sijtsma and Hemker (1998) showed that Equation 5 and Equation 6 do not imply IIO. They also showed that the monotonicity of an ISRF as in Equation 6 is neither necessary nor sufficient for IIO but, as we will see shortly, rather it is the exact definition of the ISRFs (as in the rating scale model; Andrich, 1978) or the mutual relationship between the ISRFs (see Equation 7 below) which determines whether IIO holds. Polytomous IRT models were not proposed with an eye toward the IIO property, and the dominant perspective of psychometrics on the ISRF rather than the IRF means that polytomous IRT models may or may not possess the IIO property. This is more the result of coincidence than design.

For the class of CPMs, Sijtsma and Hemker (1998) defined an order restriction on the ISRFs of the k items in the test that describe the response

probability for the same item score x, $C_{x_i}(\theta)$, $i = 1, \ldots, k$, such that

$$C_{x_1}(\theta) \leq C_{x_2}(\theta) \leq \ldots \leq C_{x_k}(\theta), \tag{7}$$

for $x = 1, \ldots, m$, and for all θ, and showed that this order restriction implies IIO. Scheiblechner (1995, Definition; also, 2003) discussed *weak item idependence*, which is an item ordering property resembling Equation 7 but without reference to a latent variable. We apply restrictions like those in Equation 7 to the classes of ACMs and CRMs. This results in IRT models that imply IIO. We show that the three general polytomous IRT models have a hierarchical relationship, and derive observable consequences, which are used to investigate in real data whether a set of k items has IIO. Next, we impose an order restriction similar to Equation 7 on the typical ISRFs from each of the three classes.

6.3 Latent Scales for Polytomous Items

For dichotomously scored items, Rosenbaum (1987a) defined a latent scale as a model for which local independence (i.e., Equation 1) holds and in each item pair (i, j) $(i < j)$ item i is uniformly more difficult than item j (Rosenbaum, 1987b), so that

$$P(X_i = 1|\theta) \leq P(X_j = 1|\theta),$$

for all θ. We generalize the concept of a latent scale to the three classes of polytomous IRT models. Using the acronym LS for latent scale, the resulting models are denoted LS-CPM, LS-CRM, and LS-ACM. We assume local independence to hold for the k polytomously scored items in the test. For scores $x = 1, \ldots, m$ on items i and j $(i < j)$, an LS-CPM is defined as

$$C_{x_i}(\theta) \leq C_{x_j}(\theta), \tag{8}$$

for all θ (equivalent to Equation 7); an LS-CRM as

$$M_{x_i}(\theta) \leq M_{x_j}(\theta), \tag{9}$$

for all θ; and an LS-ACM as

$$A_{x_i}(\theta) \leq A_{x_j}(\theta), \tag{10}$$

for all θ. Equations 8, 9, and 10 do not restrict the ordering of the ISRFs corresponding to different score categories. Equation 8 is equivalent to Equation 7 and thus implies IIO. The latent scales do not assume monotonicity. We prove the next theorem.

Theorem 1. *The three latent-scale IRT models for polytomously scored items, the LS-ACM, the LS-CRM, and the the LS-CPM have a hierarchical relation. The least restrictive of these models, the LS-CPM, implies IIO. These relationships are represented in the next scheme of logical implications:*

$$\textit{LS-ACM} \;\Rightarrow\; \textit{LS-CRM} \;\Rightarrow\; \textit{LS-CPM} \;\Rightarrow\; \textit{IIO}.$$

We prove three lemmas, which together prove Theorem 1.

Lemma 1. *The LS-ACM implies the LS-CRM.*

Proof. First note that, for $z > x$,

$$\begin{aligned}
A_{y_i}(\theta) \leq A_{y_j}(\theta) \quad &\Leftrightarrow \quad \frac{1 - A_{y_i}(\theta)}{A_{y_i}(\theta)} \geq \frac{1 - A_{y_j}(\theta)}{A_{y_j}(\theta)} \\
&\Leftrightarrow \quad \frac{P(X_i = y-1|\theta)}{P(X_i = y|\theta)} \geq \frac{P(X_j = y-1|\theta)}{P(X_j = y|\theta)} \\
&\Rightarrow \quad \prod_{y=x+1}^{z} \frac{P(X_i = y-1|\theta)}{P(X_i = y|\theta)} \geq \prod_{y=x+1}^{z} \frac{P(X_j = y-1|\theta)}{P(X_j = y|\theta)} \\
&\Leftrightarrow \quad \frac{P(X_i = x|\theta)}{P(X_i = z|\theta)} \geq \frac{P(X_j = x|\theta)}{P(X_j = z|\theta)} \\
&\Leftrightarrow \quad \frac{P(X_i = z|\theta)}{P(X_i = x|\theta)} \leq \frac{P(X_j = z|\theta)}{P(X_j = x|\theta)}.
\end{aligned} \tag{11}$$

Thus, we have shown that LS-ACM (Equation 10) implies Equation 11. Summing both its sides over $z = x+1, x+2, \ldots, m$ gives

$$\frac{P(X_i > x|\theta)}{P(X_i = x|\theta)} \leq \frac{P(X_j > x|\theta)}{P(X_j = x|\theta)},$$

which implies

$$\frac{P(X_i = x|\theta)}{P(X_i > x|\theta)} \geq \frac{P(X_j = x|\theta)}{P(X_j > x|\theta)},$$

and so

$$\begin{aligned}
\frac{P(X_i \geq x|\theta)}{P(X_i > x|\theta)} &= \frac{P(X_i = x|\theta) + P(X_i > x|\theta)}{P(X_i > x|\theta)} \\
&= \frac{P(X_i = x|\theta)}{P(X_i > x|\theta)} + 1 \\
&\geq \frac{P(X_j = x|\theta)}{P(X_j > x|\theta)} + 1 \\
&= \frac{P(X_j \geq x|\theta)}{P(X_j > x|\theta)}. \qquad (12)
\end{aligned}$$

The left and right hand sides of Equation 12 are the reciprocals of $M_{x+1_i}(\theta)$ and $M_{x+1_j}(\theta)$, respectively, so we have shown that LS-ACM implies

$$M_{x_i}(\theta) \leq M_{x_j}(\theta)$$

for all x, θ and $i < j$; that is, that LS-CRM holds. □

The following example shows that the reverse relationship between the two latent scales does not hold; that is, the LS-CRM does not imply the LS-ACM. For trichotomously scored items i and j, for some arbitrary value θ_0 let $P(X_i = \mathbf{x}|\theta_0)$ be $(\frac{1}{2}, \frac{1}{4}, \frac{1}{4})$, and let $P(X_j = \mathbf{x}|\theta_0)$ be $(\frac{1}{3}, \frac{1}{12}, \frac{7}{12})$. It may be verified that $M_{1_i} = M_{2_i} = \frac{1}{2}$, $M_{1_j} = \frac{2}{3}$, and $M_{2_j} = \frac{7}{8}$, and further that for $x = 1, 2$ it holds that $M_{x_i} < M_{x_j}$. Additional computations show that $A_{1_i} = \frac{1}{3}$, $A_{2_i} = \frac{1}{2}$, $A_{1_j} = \frac{1}{5}$, and $A_{2_j} = \frac{7}{8}$. Because $A_{1_i} > A_{1_j}$ contradicts Equation 10, the LS-ACM does not hold.

Lemma 2. *The LS-CRM implies LS-CPMs.*

Proof. We assume that the LS-CRM holds; that is, Equation 9 holds for all x

and all θ. It may be noted that

$$\begin{aligned} C_{x_i}(\theta) &= \frac{P(X_i \geq x|\theta)}{P(X_i \geq 0|\theta)} \\ &= \frac{P(X_i \geq 1|\theta)}{P(X_i \geq 0|\theta)} \times \frac{P(X_i \geq 2|\theta)}{P(X_i \geq 1|\theta)} \times \ldots \times \frac{P(X_i \geq x|\theta)}{P(X_i \geq x-1|\theta)} \\ &= \prod_{u=1}^{x} M_{u_i}(\theta). \end{aligned} \tag{13}$$

Because $M_{x_i}(\theta) \leq M_{x_j}(\theta)$ for all x and all θ, it follows from Equation 13 that

$$C_{x_i}(\theta) \leq C_{x_j}(\theta);$$

that is, that LS-CPM holds. □

The following example shows that the reverse of the implication does not hold; that is, the LS-CPM does not imply the LS-CRM. For trichotomously scored items i and j, for some arbitrary value θ_0, let $P(X_i = \mathbf{x}|\theta_0)$ be $(\frac{1}{2}, \frac{1}{4}, \frac{1}{4})$, and let $P(X_j = \mathbf{x}|\theta_0)$ be $(\frac{1}{3}, \frac{9}{24}, \frac{7}{24})$. It may be verified that $C_{1_i} = \frac{1}{2}$, $C_{2_i} = \frac{1}{4}$, $C_{1_j} = \frac{2}{3}$, and $C_{2_j} = \frac{7}{24}$. Next, it can be verified that $C_{x_i} < C_{x_j}$ for $x = 1, 2$. Finally, we find that $M_{1_i} = M_{2_i} = \frac{1}{2}$, $M_{1_j} = \frac{2}{3}$, and $M_{2_j} = \frac{7}{16}$. Because $M_{2_i} > M_{2_j}$ contradicts Equation 9, the LS-CRM does not hold.

Lemma 3. *The LS-CPM implies IIO.*

Proof. See Sijtsma and Hemker (1998), who show that Equation 8 is a sufficient (but not a necessary) condition for IIO. □

The three latent scales provide different definitions of agreement among the subjects with respect to the ordering of the items on latent variable θ, expressed by Equation 8, Equation 9, and Equation 10; See Douglas, Feinberg, Lee, Sampson, and Whitaker (1991) for related work in the context of contingency tables for ordered variables. A fourth latent scale may be defined by the combination of local independence and IIO. Theorem 1 shows that these four definitions become progressively weaker, going from the LS-ACM via the LS-CRM and the LS-CPM to IIO. Thus far, in psychometrics item orderings have

been defined in terms of expected item scores such as IIO. IIO is the weakest form of agreement among the respondents with respect to the ordering of the items on latent variable θ. However, given their relationships to particular task structures (Van Engelenburg, 1997) the other latent scales are also possible ways of defining this agreement.

The task types Van Engelenburg (1997) suggested for the three classes of polytomous IRT models produce item scores, the structure of which also matches the formal structure of the different latent scales; that is, tasks best suited for CPM models are best suited for LS-CPM models, and so on. The relationships between task structure and latent scale are not logically compelling. However, the structure of the task may give direction for the choice of an appropriate latent scale for analyzing one's data.

6.4 Manifest Properties of Latent Scales

In this section, we derive three observable consequences or manifest properties from the latent scales. In particular, the LS-ACM implies the *increasingness in transposition* (IT) property (Theorem 3); the LS-CPM implies the *manifest scale cumulative probability model* (MS-CPM) property (Theorem 2); and IIO implies the *manifest invariant item ordering* (MIIO) property (Corollary). Latent scales and observable properties are related as:

$$\begin{array}{ccccccc} \text{LS-ACM} & \Rightarrow & \text{LS-CRM} & \Rightarrow & \text{LS-CPM} & \Rightarrow & \text{IIO} \\ \Downarrow & & & & \Downarrow & & \Downarrow \\ \text{IT} & & & & \text{MS-CPM} & & \text{MIIO} \end{array}$$

The manifest properties can be used as a basis for investigating whether support can be found in the data for a particular latent scale. In the next section, to this end we discuss the IT method, the MS-CPM method, and the MIIO method. As with all model-data fit research, fit is necessarily assessed using observable consequences, which can only provide incomplete information about a model. Hence, conclusions should always be drawn with caution. Next, we prove the downward implications in the scheme.

6.4.1 Manifest Scales

Let Y be a manifest variable with realization y that is independent of item scores X_i and X_j given θ. For example, Y may be a function of the $k-2$ items in the test without the items i and j, the sum score obtained on a different test, or an indicator of group membership. Replacing the latent variable θ in the latent scales defined in Equations 8, 9, and 10 by the manifest variable Y yields their manifest scale (MS) analogues. For example, for $i < j$ and for $x = 1, \ldots, m$, an MS-CPM is defined as $C_{x_i}(Y) \leq C_{x_j}(Y)$ for all values of Y (cf. Equation 8). Similarly, for $i < j$ an MIIO is defined as $E(X_i|Y) \leq E(X_j|Y)$ for all y.

In Theorem 2, we prove that the LS-CPM implies the MS-CPM, and the Corollary shows that an IIO implies an MIIO. The LS-ACM and the LS-CRM do not imply their manifest analogues. Thus, if in empirical data analysis the MS-CPM and MIIO are satisfied, some support is found for the theoretical LS-CPM and IIO, respectively, but if the MS-ACM and the MS-CRM are satisfied, this cannot be taken as support for their latent scale analogues. Because fitting MS-ACMs and MS-CRMs do not support latent scales, they are not further pursued here.

Theorem 2. *The LS-CPM implies the MS-CPM.*

Proof. Let $F(\theta)$ be the cumulative distribution function of θ. Multiplying both sides of Equation 8 by $P(Y = y|\theta)$ and integrating over θ yields

$$\begin{aligned} C_{x_i}(\theta) \leq C_{x_j}(\theta),\ \forall\theta \quad &\Leftrightarrow \\ P(X_i \geq x|\theta) \leq P(X_j \geq x|\theta),\ \forall\theta \quad &\Rightarrow \end{aligned} \tag{14}$$

$$\int_{\theta} P(X_i \geq x|\theta)P(Y = y|\theta)dF(\theta) \leq \int_{\theta} P(X_j \geq x|\theta)P(Y = y|\theta)dF(\theta). \tag{15}$$

Because Y is conditionally independent of X_i and X_j, Equation 15 is equivalent

to

$$\begin{aligned}\int_{\theta} P(X_i \geq x, Y = y|\theta)dF(\theta) &\leq \int_{\theta} P(X_j \geq x, Y = y|\theta)dF(\theta) \quad \Leftrightarrow \\ P(X_i \geq x, Y = y) &\leq P(X_j \geq x, Y = y) \quad \Leftrightarrow \\ P(X_i \geq x|Y = y) &\leq P(X_j \geq x|Y = y). \end{aligned} \tag{16}$$

The proof holds for all x and y, and all $i < j$. □

IIO does not imply MS-CPM. For reasons of space, we do not provide a counter example. Counter examples showing that the LS-ACM and the LS-CRM do not imply their manifest analogues are complex and can be obtained from the first author. For an appropriate choice of variable Y, in real data it can be investigated whether the MS-CPM property (Equation 16) is satisfied. The next section discusses how the MS-CPM method based on Equation 16 may be used in the analysis of real data.

Corollary. IIO implies MIIO.

Proof. Theorem 2 states that for $x = 1, \ldots, m$, for $i < j$, and all θ that

$$P(X_i \geq x|\theta) \geq P(X_j \geq x|\theta) \Rightarrow P(X_i \geq x|Y = y) \geq P(X_j \geq x|Y = y)$$

for all y. This result can also be shown to hold for sums of cumulative response probabilities. For $x = 1, \ldots, m$, for $i < j$, and all θ

$$\begin{aligned}\sum_{x=1}^{m} P(X_i \geq x|\theta) \geq \sum_{x=1}^{m} P(X_j \geq x|\theta) \Rightarrow \\ \sum_{x=1}^{m} P(X_i \geq x|Y = y) \geq \sum_{x=1}^{m} P(X_j \geq x|Y = y)\end{aligned}$$

for all y. This implication is equivalent to

$$E(X_i|\theta) \geq E(X_j|\theta) \Rightarrow E(X_i|Y) \geq E(X_j|Y); \tag{17}$$

also see Shaked and Shantikumar (1994, p. 4). □

For an appropriate choice of variable Y, the MIIO property (i.e., the right-hand side of Equation 17) can be investigated in real data so as to collect support in favor of IIO. The corresponding MIIO method is discussed in the next section. Because the most general latent scale, which is the IIO, implies MIIO, by implication the previous three latent scales in the ordered series also imply MIIO. In the same vein, because the LS-CPM implies the MS-CPM, the preceding and most restrictive latent scales, the LS-ACM and the LS-CRM, also imply the MS-CPM.

6.4.2 Increasingness in Transposition

Rosenbaum (1987a) used the manifest IT property (Hollander, Proschan, & Sethuraman, 1977) to investigate whether a set of dichotomously scored items forms a latent scale. We adapt the results presented by Rosenbaum (1987a) to investigate whether a set of polytomously scored items constitute a latent scale (Equations 8, 9, and 10). First, we introduce some notation.

The set of items and their indices $\mathcal{T}$ is divided into two subsets $(\mathcal{S}, \mathcal{R})$. Subset $\mathcal{S}$ contains at least two items, and subset $\mathcal{R}$ contains the remaining items. The scores on the items in $\mathcal{S}$ are collected in item-score vector $\mathbf{x}_{\mathcal{S}}$, and the scores on the items in $\mathcal{R}$ in item-score vector $\mathbf{x}_{\mathcal{R}}$. We define item difficulty as the expected score on an item across the distribution of θ, denoted $F(\theta)$: that is, $E(X_i) = \int E(X_i|\theta)dF(\theta)$, for $i = 1, \ldots, k$. Let i and j be two items from $\mathcal{S}$, and let $i < j$ denote that item i is at least as difficult as item j; that is, $E(X_i) \leq E(X_j)$. Then, $x_i > x_j$ means that the score on the more difficult item i is higher than the score on the easier item j. Furthermore, let $h(\mathbf{X}_{\mathcal{R}})$ be a function of the scores on the items in $\mathcal{R}$. For example, $h(\mathbf{X}_{\mathcal{R}})$ may be the sum score on the items in $\mathcal{R}$, or it may be a single item score.

Vector $\mathbf{x}'_{\mathcal{S}}$ is defined as a *transposition* of vector $\mathbf{x}_{\mathcal{S}}$, if one or more reversals of two scores in vector $\mathbf{x}_{\mathcal{S}}$ produce vector $\mathbf{x}'_{\mathcal{S}}$ (Hollander et al., 1977). For example, $\mathbf{x}'_{\mathcal{S}} = (1, 1, 0, 2)$ is a transposition of $\mathbf{x}_{\mathcal{S}} = (1, 2, 0, 1)$, because the reversal

of x_2 and x_4 in $\mathbf{x}_\mathcal{S}$ produces $\mathbf{x}'_\mathcal{S}$. Also, two reversals are needed to go from $\mathbf{x}_\mathcal{S}$ to $\mathbf{x}''_\mathcal{S} = (0, 1, 1, 2)$. Finally, $\mathbf{x}'''_\mathcal{S} = (1, 2, 1, 2)$ and $\mathbf{x}_\mathcal{S}$ are not transpositions of one another.

Next, we consider two vectors $\mathbf{x}_\mathcal{S}$ and $\mathbf{x}'_\mathcal{S}$, which are transpositions of one another, and define the *partial order* '$\prec$' on these vectors. A partial order $\mathbf{x}_\mathcal{S} \prec \mathbf{x}'_\mathcal{S}$ means that $\mathbf{x}_\mathcal{S}$ produces $\mathbf{x}'_\mathcal{S}$ when interchanging item scores in $\mathbf{x}_\mathcal{S}$ implies that higher item scores are moved to the right while lower item scores are moved to the left. In the previous example, $\mathbf{x}_\mathcal{S}$ produced $\mathbf{x}'_\mathcal{S}$ when the higher score $x_2 = 2$ was interchanged with the lower score $x_4 = 1$. What happens is that, given the item ordering $E(X_1) \leq E(X_2) \leq E(X_3) \leq E(X_4)$, the ordering of item scores in the resulting vector $\mathbf{x}'_\mathcal{S}$ better matches the item ordering by difficulty than in the original vector $\mathbf{x}_\mathcal{S}$.

Let $P\big(\mathbf{x}_\mathcal{S}|h(\mathbf{X}_\mathcal{R})\big)$ be the probability of item-score vector $\mathbf{x}_\mathcal{S}$ conditional on score function $h(\mathbf{X}_\mathcal{R})$. Under some IRT models, such probabilities can be ordered in $\mathbf{X}_\mathcal{S}$ (i.e., for different vectors $\mathbf{x}_\mathcal{S}$) provided the item-score vectors are partially ordered. More specifically, conditional on function $h(\mathbf{X}_\mathcal{R})$, the probabilities of two vectors $\mathbf{x}_\mathcal{S}$ and $\mathbf{x}'_\mathcal{S}$, which are partially ordered by $\mathbf{x}_\mathcal{S} \prec \mathbf{x}'_\mathcal{S}$, are ordered such that $P\big(\mathbf{x}_\mathcal{S}|h(\mathbf{X}_\mathcal{R})\big) \leq P\big(\mathbf{x}'_\mathcal{S}|h(\mathbf{X}_\mathcal{R})\big)$. When such an ordering is possible, the probabilities are *increasing in transposition* in $\mathbf{X}_\mathcal{S}$. Suppose, the partially ordered vectors $\mathbf{x}_\mathcal{S}$ and $\mathbf{x}'_\mathcal{S}$ differ with respect to two or more transpositions; then, successive transpositions step-by-step move higher scores from $\mathbf{x}_\mathcal{S}$ to the right until $\mathbf{x}'_\mathcal{S}$ is obtained. Vectors $\mathbf{x}_\mathcal{S}$ and $\mathbf{x}'_\mathcal{S}$ and the vectors obtained in each step moving from $\mathbf{x}_\mathcal{S}$ to $\mathbf{x}'_\mathcal{S}$ are collected in a set denoted $\mathcal{V}$. It may be noted that set $\mathcal{V}$ contains only those vector permutations that are partially ordered. Then, the formal definition of functions that are IT in $\mathbf{X}_\mathcal{S}$ is the following: $P(.)$ is IT in $\mathbf{X}_\mathcal{S}$ for function $h(.)$ if for all $\{\mathbf{x}_\mathcal{S}, \mathbf{x}'_\mathcal{S}\} \in \mathcal{V}$, which have a partial ordering $\mathbf{x}_\mathcal{S} \prec \mathbf{x}'_\mathcal{S}$, we have that

$$P\big(\mathbf{x}_\mathcal{S}|h(\mathbf{X}_\mathcal{R})\big) \leq P\big(\mathbf{x}'_\mathcal{S}|h(\mathbf{X}_\mathcal{R})\big).$$

As an example, for the sake of simplicity we assume that $\mathcal{R} = \emptyset$. Thus,

$\mathcal{S} = \mathcal{T}$, so that $P\big(\mathbf{x}_{\mathcal{S}}|h(\mathbf{X}_{\mathcal{R}})\big) = P(\mathbf{x}_{\mathcal{S}})$. Now, because the vectors $(2,1,1,0)$ and $(0,1,1,2)$ are partially ordered, the IT property implies that $P(2,1,1,0) \leq P(0,1,1,2)$.

Theorem 3. *The LS-ACM implies IT.*

Proof. The point of departure is Equation 11, which holds under the LS-ACM. For $0 \leq y < z \leq m$ and $i < j$, Equation 11 is equivalent to

$$\frac{P(X_i = z|\theta)P(X_j = y|\theta)}{P(X_i = y|\theta)P(X_j = z|\theta)} \leq 1. \tag{18}$$

For dichotomously scored items with $y = 0$ and $z = 1$, Rosenbaum (1987a, Theorem 1) showed that Equation 18 implies IT. We extend Rosenbaum's proof to polytomous items.

Let k_s be the number of items in subset $\mathcal{S}$, and let $\mathcal{S} \setminus \{i,j\}$ denote the subset of $k_s - 2$ items that remain in $\mathcal{S}$ after items i and j have been excluded. For subset $\mathcal{S}$ including items i and j (i.e., $\{i,j\} \in \mathcal{S}$), Equation 18 is equivalent to

$$\frac{P(X_i = z|\theta)P(X_j = y|\theta)}{P(X_i = y|\theta)P(X_j = z|\theta)} \prod_{u \in \mathcal{S}\setminus\{i,j\}} \frac{P(X_u = x_u|\theta)}{P(X_u = x_u|\theta)} \leq 1. \tag{19}$$

Because of local independence (Equation 1), Equation 19 can be written as

$$\frac{P(X_1 = x_1, \ldots X_i = z, \ldots, X_j = y, \ldots, X_{k_s} = x_{k_s}|\theta)}{P(X_1 = x_1, \ldots, X_i = y, \ldots, X_j = z, \ldots, X_{k_s} = x_{k_s}|\theta)} \leq 1. \tag{20}$$

The item-score vector in the numerator is denoted by $\mathbf{x}_{\mathcal{S}}$ and the item-score vector in the denominator by $\mathbf{x}'_{\mathcal{S}}$. It may be noted that $\mathbf{x}_{\mathcal{S}}$ and $\mathbf{x}'_{\mathcal{S}}$ are partially ordered, $\mathbf{x}_{\mathcal{S}} \prec \mathbf{x}'_{\mathcal{S}}$. We rewrite Equation 20 as

$$\frac{P(\mathbf{x}_{\mathcal{S}}|\theta)}{P(\mathbf{x}_{\mathcal{S}'}|\theta)} \leq 1. \tag{21}$$

Hollander et al. (1977, Theorem 3.2) show that Equation 21 implies

$$\int \frac{P(\mathbf{x}_{\mathcal{S}}|\theta)}{P(\mathbf{x}'_{\mathcal{S}}|\theta)} dF(\theta) \leq 1. \tag{22}$$

Finally, Equation 22 implies the manifest IT property,

$$\frac{P\big(\mathbf{x}_{\mathcal{S}}|h(\mathbf{X}_{\mathcal{R}})\big)}{P\big(\mathbf{x}'_{\mathcal{S}}|h(\mathbf{X}_{\mathcal{R}})\big)} \leq 1 \Leftrightarrow P\big(\mathbf{x}_{\mathcal{S}}|h(\mathbf{X}_{\mathcal{R}})\big) \leq P\big(\mathbf{x}'_{\mathcal{S}}|h(\mathbf{X}_{\mathcal{R}})\big) \tag{23}$$

(cf. Rosenbaum, 1987a, Theorem 1). □

The other latent scales (LS-CRM, LS-CPM) and IIO do not imply IT. Counter examples may be obtained from the first author. It may be noted that Equation 23 holds for any conditioning variable Y that is independent of the items in $\mathcal{S}$, but that we used $h(\mathbf{X}_{\mathcal{R}})$ to stay close to previous work (Rosenbaum, 1987a; Sijtsma & Junker, 1996). The proof may be extended to partially ordered vectors $\mathbf{x}_{\mathcal{S}}$ and $\mathbf{x}'_{\mathcal{S}}$ that differ with respect to two or more transpositions following a step-by-step permutation of $\mathbf{x}_{\mathcal{S}}$ into $\mathbf{x}'_{\mathcal{S}}$ by successive transpositions that move higher scores to the right, and applying Equation 20 successively. In the next section, we discuss how the IT method based on the IT property can be used in the analysis of real data to collect support in favor of the LS-ACM.

6.5 Methods for Data Analysis

For realistic numbers of items, the investigation of the MIIO, MS-CPM, and IT properties produces multiple results, which have to be combined for each property to decide whether that property holds in the data and, hence, provides support for a particular latent scale. Ligtvoet et al. (in press) proposed a method for dealing with multiple results when testing the MIIO property. Here, we adapt this method to the MS-CPM and IT properties, but first we explain the MIIO method (see Ligtvoet et al., in press, for details).

For each item-pair (i, j) $(i < j)$, it is investigated whether it violates MIIO (Equation 17). This produces $\frac{1}{2} \times k \times (k-1)$ Boolean outcomes on violation of MIIO. The statistical testing procedure for one item-pair (i, j) is as follows. Variable Y in Equation 17 is replaced by rest score $R_{ij} = \sum_{i' \neq i,j} X_{i'}$, so that MIIO is investigated by checking whether $E(X_i|R_{ij} = r) \leq E(X_j|R_{ij} = r)$ for

$r = 0, \ldots, k-2$. If the sample means are reversely ordered (i.e., $\overline{X}_i|R_{ij} = r > \overline{X}_j|R_{ij} = r$), a one-sided t-test is used for deciding whether the violation is significant. To avoid testing violations that are too small on a scale ranging from 0 to m to be of practical interest, violations smaller than $m \times .03$ are ignored. Adjacent restscore groups $r, r+1, l \ldots$ containing few observations may be joined to gain statistical power (Molenaar & Sijtsma, 2000, p. 67). If one or more t-tests of violations in excess of $m \times .03$ are significant, the item pair violates MIIO.

If MIIO does not hold for all k items, items are removed one-by-one until a subset remains for which MIIO holds (Ligtvoet et al., in press). A backward item-selection procedure reaches this goal while removing as few items as possible. This is done in the first step by counting for each item how many of the $k-1$ item pairs in which the item is involved violate MIIO significantly according to the t-test procedure. The item with the highest count is removed first; for the remaining $k-1$ items the counts are redone without the item that was removed, and if there are item pairs violating MIIO, the item having the highest count is removed; and this procedure is repeated until there are no item-pairs left that violate MIIO. If two or more items have the highest count, then the item that has the lowest scalability value is removed (Ligtvoet et al., in press). The same rest score based on $k-2$ items is used throughout so as to minimize the risk of chance capitalization. We adapt this strategy to the MS-CPM (Equation 16) and IT (Equation 23) properties, thus producing the MS-CPM and IT methods.

For the MS-CPM property, let $P(X_i \geq x|Y = y)$ and $P(X_j \geq x|Y = y)$ in Equation 16 be denoted *pair of manifest ISRFs* (i, j, x). For each pair of manifest ISRFs (rather than for each item pair) it is investigated whether the pair violates the MS-CPM property, which produces $\frac{1}{2} \times k \times (k-1) \times m$ Boolean outcomes. The testing procedure for one pair of manifest ISRFs (i, j, x) is as follows. Just as for MIIO, variable Y in Equation 16 is replaced by rest score R_{ij}. Hence, the MS-CPM property is investigated by checking whether $P(X_i \geq x|R_{ij} = r) \leq P(X_j \geq x|R_{ij} = r)$ for all r. If the sample fractions

are reversely ordered (i.e., $\hat{P}(X_i \geq x|R_{ij} = r) > \hat{P}(X_j \geq x|R_{ij} = r)$, a z-test (Molenaar & Sijtsma, 2000, p. 78) is used to decide whether the violation is significant . Following recommendations by Molenaar and Sijtsma (2000, pp. 67-70), violations smaller than .03 are ignored.

For the IT property (Equation 23), the method is adapted as follows. We consider item pairs, so that item-score vectors $\mathbf{x}_{\mathcal{S}}$ and $\mathbf{x}_{\mathcal{S}'}$ in Equation 23 are reduced to two elements: $\mathbf{x}_{\mathcal{S}} = (u, v)$ and $\mathbf{x}_{\mathcal{S}'} = (v, u)$ $(u = 0, \ldots, m-1; v = u+1, \ldots, m)$. Consistent with MIIO and MS-CPM, function $h(\mathbf{X}_{\mathcal{R}}) = R_{ij}$. Let $P\big(\mathbf{X}_{\mathcal{S}} = \mathbf{x}_{\mathcal{S}}|h(\mathbf{X})\big) = P(X_i = u, X_j = v|R_{ij})$ and $P\big(\mathbf{X}_{\mathcal{S}'} = \mathbf{x}_{\mathcal{S}'}|h(\mathbf{X})\big) = P(X_i = v, X_j = u|R_{ij})$ be the *pair of bivariate conditional probabilities* (i, j, u, v). For each pair of bivariate conditional probabilities, it is investigated whether the pair violates the IT property. This produces $\frac{1}{2} \times k \times (k-1) \times \frac{1}{2} \times m \times (m-1)$ Boolean outcomes. The testing procedure for pair (i, j, u, v) is as follows. IT is investigated by checking whether $P(X_i = v, X_j = u|R_{ij} = r) \leq P(X_i = u, X_j = v|R_{ij} = r)$. If the sample fractions are reversely ordered (i.e., $\hat{P}(X_i = v, X_j = u|R_{ij} = r) > \hat{P}(X_i = u, X_j = v|R_{ij} = r)$), the McNemar (1947) test is used to decide whether the violation is significant. Let $n_{uv|r}$ and $n_{vu|r}$ denote the sample sizes of the relevant fractions, then under the null-hypothesis that the two bivariate conditional probabilities are equal,

$$X^2 = \frac{(n_{uv|r} - n_{vu|r})^2}{n_{uv|r} + n_{vu|r}}$$

has an asymptotic chi-square distribution with 1 degree of freedom. As with MS-CPM, violations smaller than .03 are ignored.

For the MS-CPM and IT methods, the backward item-selection procedures are formally identical to that of the MIIO method, and are not repeated here. For confirmatory results from the MS-CPM method, we infer that the LS-CPM supports the final item subset, and for confirmatory results from the IT method, we infer that the LS-ACM supports the final item subset. Many different strategies for testing the IT property are possible (see Sijtsma & Junker, 1996, p. 90) but they are beyond the scope of this study.

Table 6.1: Violations of MIIO, MS-CPM, and IT for Coping-Strategy Data.

Item	Mean	MIIO		MS-CPM		IT	
1. Call environmental agency	.264	0	0	2	0	1	NA
2. File a complaint with producer	.353	0	0	1	0	0	0
3. Go elsewhere for fresh air	.535	0	0	4	NA	NA	NA
4. Experience unrest	.651	0	0	2	0	1	0
5. Try to find solutions	.818	2	NA	NA	NA	NA	NA
6. Do something to get rid of it	.860	1	0	3	NA	NA	NA
7. Talk to friends and family	.983	1	0	2	0	1	0
8. Search source of malodor	1.849	0	0	0	0	0	0

Note: NA = Not available.

6.6 Real-Data Examples

We discuss two real-data examples to illustrate how the MIIO, MS-CPM, and IT methods may be used to investigate the latent scales. We used the function `check.iio` from the R package `mokken` (Van der Ark, 2007).

6.6.1 Ordering coping strategies

Data came from eight polytomously scored items administered to 828 subjects (Cavalini, 1992) asking them how they coped actively with the bad smell from a factory in the neighborhood of their homes. Table 6.1 shows the items ordered and numbered by increasing item mean. Items have four ordered answer categories, "never" (score 0), "'seldom" (1), "often" (2), and "always" (3) (i.e., $m = 3$). The items constitute an ordinal scale according to the monotone homogeneity model (Sijtsma & Molenaar, 2002, chap. 3). The items require global assessment using a rating scale (Van Engelenburg, 1997); hence, the LS-CPM may be the appropriate model to analyze the data. The aim of the analysis was to select a subset of items that constitute an LS-CPM scale, and represent a set of invariantly ordered coping reactions.

First, we tested the data for MIIO (Equation 17), which is the least restrictive manifest ordering property, to identify items grossly violating an invariant ordering. We found that two out of the $\frac{1}{2} \times 8 \times 7 = 28$ item pairs (i.e., item-pairs

(5,6) and (5,7)) violated MIIO. Table 6.1 (fourth column) shows that item 5 was included in two item pairs violating MIIO, and items 6 and 7 were each included in one item pair. Removal of item 5 from the 8-item set (Table 6.1, fifth column) resulted in a 7-item set without violations, which provided support for IIO.

Next, we tested the MS-CPM (Equation 16) for the remaining 7 items, and found that seven out of the $\frac{1}{2} \times 7 \times 6 \times 3 = 63$ pairs of manifest ISRFs showed violations (Table 6.1; sixth column). Items 3 and 6 together were involved in all violations. Removal of these items from the 7-item set (Table 6.1, column 7) resulted in a 5-item scale for which the MS-CPM could not be rejected, and which provided support for the LS-CPM.

For the purpose of illustration, we also investigated the IT property (Equation 23) for the remaining 5 items. Two out of the $\frac{1}{2} \times 5 \times 4 \times \frac{1}{2} \times 4 \times 3 = 60$ pairs of bivariate conditional probabilities violated IT; that is, $\hat{P}(X_1 = 2, X_4 = 3|R_{ij} \in \{6, \ldots, 9\}) > \hat{P}(X_1 = 3, X_4 = 2|R_{ij} \in \{6, \ldots, 9\})$ and $\hat{P}(X_1 = 2, X_7 = 3|R_{ij} \in \{6, \ldots, 9\}) > \hat{P}(X_1 = 3, X_7 = 2|R_{ij} \in \{6, \ldots, 9\})$. Both were significant. Removal of item 1 from the 5-item set (Table 6.1, seventh column) resulted in a scale without violations, thus providing support for the LS-ACM.

6.6.2 Dutch History

The data were scores on three items collected from 752 students. The items were selected from a 40-item exam on Dutch history to illustrate the LS-ACM rather than the LS-CPM used in the first example. In each of the items, four historical events are presented and the student is asked whether the first event preceded the second, the second the third, and the third the fourth. The remaining 37 items had different item formats and could not be used for illustrating the LS-ACM.

The events of zero or one correct answer were relatively rare. Hence, the three items were scored as follows: 0 for zero or one correct answer; 1 for two correct answers; and 2 for three correct answers. Items were numbered following their ascending sample means, $\overline{X}_1 = 1.243$, $\overline{X}_2 = 1.327$, and $\overline{X}_3 = 1.386$. The

task structure suggests that the subtasks may be solved in an arbitrary order (Van Engelenburg, 1997). Thus, the LS-ACM may be the appropriate model for investigating the item ordering. With only three items, item selection is not of interest here. It may be noted that for three items, the rest score is the score on one item only.

The LS-ACM was investigated checking $\frac{1}{2} \times 3 \times 2 \times \frac{1}{2} \times 3 \times 2 = 9$ pairs of bivariate conditional probabilities (Equation 23), one of which was significant: $\hat{P}(X_2 = 1, X_3 = 0|X_1 = 0) > \hat{P}(X_2 = 0, X_3 = 1|X_1 = 0)$: $X^2 = 3.90$, $df = 2$, $p = .048$. Based on the IT method, the LS-ACM should be rejected. We also used the MS-CPM and MIIO methods for data analysis. For MS-CPM, we found that two out of the six pairs of manifest ISRFs violated the MS-CPM. For MIIO we did not find violations.

6.7 Summary and Discussion

Several applications of tests are based on the assumption that the ordering of the items by difficulty is the same for every subject to which the test is administered, a property known as invariant item ordering. In this study, the concept of IIO for a set of polytomous items was extended to three classes of polytomous IRT models, which are the latent-scale adjacent category models, the latent-scale continuation ratio models, and the latent-scale cumulative probability models. It was proven that these latent scales are hierarchically ordered, and that each implies the IIO property using progressively weaker forms of agreement among respondents with respect to the ordering of the items on latent variable θ.

The significance of this result was that it enabled the derivation of three observable consequences. The latent-scale adjacent category model implies the property of increasingness in transposition in partially ordered item-score vectors. The presence of this structure in the data supports the latent-scale adjacent category model, which implies IIO. The latent-scale cumulative probability model implies the manifest-scale cumulative probability model (MC-

CPM) property, of which the presence in the data supports the model and, consequently, IIO. Finally, the IIO property implies the manifest IIO property, which can be used to find direct support for IIO without the intervention of a polytomous IRT model. If the item type used matches the structure of one of the latent scales, it may be preferable to first investigate whether this model fits the data. Unfortunately, so far we have not been able to derive observable consequences for the latent-scale continuation ratio model. This is a target for future research.

Simply checking the MIIO, MS-CPM, and IT properties in data collected from realistic test lengths produces much detailed results, which complicates drawing inferences about the fit of a latent scale. Hence, we adapted a methodology for investigating MIIO (Ligtvoet et al., in press), which reduces large numbers of detailed results to one final outcome, to be used also for investigating MS-CPM and IT. For example, for 20 items with five answer categories, the IT method produces 900 series of local tests and, depending on the number of rest scores left after joining small-frequency scores, each typically contains four or five tests. We only count for each item the number of violations, and use these counts in the backward item-selection algorithm.

Rather than using local tests, one may prefer a global goodness-of-fit statistic, which assesses all violations simultaneously (but see Molenaar, 2004, who warns against the lack of diagnostic information provided by such global test results). Marginal models are a viable approach for simultaneous testing (e.g., Van der Ark, Croon, & Sijtsma, 2008). We ignored small violations of the MIIO, MS-CPM, and IT properties but did not adjust the nominal Type-I error-rate for multiple testing, which is consistent with model-fit investigation in nonparametric IRT (Sijtsma & Molenaar, 2002). More research is needed to find the proper balance between pre-selecting ignorable sample violations and an adequate Type I error. Different choices can be made with respect to conditioning variable Y. For the investigation of IT, the number of items investigated simultaneously may be varied.

Only few studies have addressed the ordering of polytomously scored items,

let alone IIO (for exceptions, see Ligtvoet et al., in press; Sijtsma & Hemker, 1998). This is remarkable because many test applications assume that the item ordering is the same for all subjects. However, as a rule this is not empirically established, and it seems that often it is not realized that IIO is a strong property which cannot be assumed to hold just like that. This study provides a step in the direction of the development of a sound psychometric theory for latent scales and IIO of polytomously scored items, and of data-analysis methods that can be used for investigating whether a latent scale or IIO holds.

Epilogue

IIO is a highly underrated property of a test or questionnaire. Both test theorists and practitioners assume all too easy that if an item is more difficult than another item for a particular subject that this difficulty ordering also holds for all other subject from the population of interest. This thesis not only shows that this is indeed a strong assumption that is not easily satisfied for a particular test or questionnaire in a particular group but also that the investigation of IIO in itself is difficult to accomplish and that many conceptual and statistical problems need to be overcome. This thesis provides a set of essays on possible avenues for IIO research but it is only a start.

Two avenues for future research with respect to IIO may be the following. First, chapter 5 showed that an excessively large sample is needed in order to distinguish realistic numbers of IRFs. If the purpose of research indeed is to distinguish the k IRFs from the items in the test of questionnaire, selecting an item subset with distinguishable and nonintersecting IRFs leads to an unacceptable loss of information, both with respect to test-score reliability and coverage of the attribute measured by the test. The solution may be too distribute the k items into clusters of items in such a way that the clusters are invariantly ordered ignoring the ordering of items within clusters.

For example, we may consider an invariant cluster ordering based on average IRFs within clusters and require that these averages do not intersect across different clusters. Wainer, Bradlow, and Wang (2007) provide suggestions for how to cluster the items into so-called item testlets. If the original k items are dichotomously scored (chapter 5), the methods for investigating IIO for

polytomously scored items (chapters 2, 3, 4 and 6) can be readily applied, with the conditional expectations taken over the item scores for those items pertaining to the same cluster. Hence, the first avenue for future IIO research may consider a courser approach to IRT by taking clusters of items as the unit of interest, rather than each individual item.

The second future avenue for IIO research corresponds to the situation where the actual ordering of items is subordinate to the global evaluation of IIO. For example, in person-fit analysis (e.g., Emons, Sijtsma, & Meijer, 2004) IIO is assumed for large numbers of items but only to allow the person response function to be fitted, for example, by means of kernel smoothing. Future research should reveal the robustness of such curve fitting to minor violations of IIO. For example, simulation studies may focus on the required *minvi* value or H^T value (chapter 2) for which the person response function decreases as the items become more difficult.

In general, the importance of IIO or the required level of agreement of subjects on the ordering of items (e.g., chapter 6) depends on the application of the test or the questionnaire. If one is only interested in the ordering of subjects by means of a test score, the monotonicity assumption (in addition to unidimensionality and local independence) may suffice (e.g., chapter 4) and IIO need not be assessed. Most applications for which a test is intended may not be directed at the item difficulties. Yet, as my role in statistics is not to find excuses for researchers not to test certain assumptions, I would encourage test developers to assess whether or not IIO holds for sets of items, even if it only serves as a standard against which the users of a test can assess whether the test actually matches the goal of their application (or method of item administration). Also, a note in a test manual concerning IIO may warn researchers that the attained ordering of items on the basis of the averaged values may not give information about the item difficulties at the lower aggregate level of individual subjects.

References

Agresti, A. (1990). *Categorical data analysis.* New York: Wiley.

Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Csáki (Eds.), *Procedings 2nd International Symposium on Information Theory*, pp. 267-281. Budapest: Akadémiai Kiadó.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*, 561-573.

Bartolucci, F. & Forcina, A. (2005). Likelihood inference on the underlying structure of IRT models. *Psychometrika, 70*, 31-43.

Birnbaum, A. (1968). Some latent trait models and their uses in inferring an examinee's ability. In F. M. Lord, & M. R. Novick, *Statistical theories of mental test scores* (pp. 397-479). Reading, MA: Addison-Wisley.

Bleichrodt, N., Drenth, P. J. D., Zaal, J. N., & Resing, W. C. M. (1987). Revisie Amsterdamse Kinder Intelligentie Test. Handleiding *[Revision Amsterdam Child Intelligence Test. Manual]*. Lisse, The Netherlands: Swets & Zeitlinger.

Bouwmeester, S., & Sijtsma, K. (2007). Latent class modeling of phase transition in the development of transitive reasoning. *Multivariate Behavioral Research, 42*, 457-480.

Cavalini, P. M. (1992). *It's an ill wind that brings no good. Studies on odour annoyance and the dispersion of odorant concentrations from industries.* Unpublished doctoral disseratation. University of Groningen, The Netherlands.

Chang, H., & Mazzeo, J. (1994). The unique correspondence of the item re-

sponse function and item category response function in polytomously scored item response models. *Psychometrika, 59*, 391-404.

Clogg, C. C. (1988). Latent class models for measuring. In R. Langeheine & J. Rost (Eds.), *Latent traits and latent class models* (pp. 173-205). New York: Plenum Press.

Coombs, C. H. (1964). *A theory of data.* New York: Wiley.

Croon, M. A. (1990). Latent class analysis with ordered latent classes. *British Journal of Mathematical and Statistical Psychology, 43*, 171-192.

Croon, M. A. (1991). Investigating Mokken scalability of dichotomous items by means of ordinal latent class analysis. *British Journal of Mathematical and Statistical Psychology, 44*, 315-331.

Douglas, R., Fienberg, S. E., Lee, M. -L. T., Sampson, A. R., & Whitaker, L. R. (1991). Positive dependence concepts for ordinal contingency tables. In H. W. Block, A. R. Sampson, & T. H. Savits (Eds.), *Topics in statistical dependence* (pp. 189-202). Hayward, CA: Institute of Mathematical Statistics.

Ellis, J. L. (1994). *Foundations of monotone latent variable models.* Unpublished doctoral dissertation, Catholic University of Nijmegen, The Netherlands.

Emons, W. H. M., Sijtsma, K., & Meijer, R. R. (2004). Testing hypotheses about the person-response function in person-fit analysis. *Multivariate Behavioral Research, 39*, 1-35.

Felling, A., Peters, J., & Schreuder, O. (1987). *Religion in Dutch society 85; documentation of a national survey on religious and secular attitudes in 1985.* Amsterdam: Steinmetz Archive.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics, 1*, 209-230.

Fischer, G. H. (1974). *Einführung in die Theorie psychologischer Tests* [Introduction to psychological test theory]. Bern, Switzerland: Huber.

Fischer, G. H. (1995). Derivations of the Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 15-38). New York: Springer.

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). "Mini-Mental State" A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research, 12*, 189-198.

Gamerman, D. (1997). *Markov chain Monte Carlo: Stochastic simulation for Bayesian inference.* London: Chapman & Hall.

Gelfand, A. E., Smith, A. F. M., & Lee, T. -M. (1992). Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling. *Journal of the American Statistical Association, 87*, 523-532.

Gelman, A., Meng, X. L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica, 6*, 733-807.

Glas, C. A. W., & Verhelst, N. D. (1995). Testing the Rasch model. In G. H. Fischer and I. W. Molenaar (Eds), *Rasch models: Foundations, recent developments, and applications* (pp. 69-96). New York: Springer.

Gough, H. G., & Heilbrun, A. B. (1980). *The Adjective Check List, Manual 1980 Edition.* Palo Alto, CA: Consulting Psychologists Press.

Grayson, D. A. (1988). Two-group classification in latent trait theory: Scores with monotone likelihood ratio. *Psychometrika, 53*, 383-392.

Heinen, T. (1996). *Discrete latent variable models.* Thousand Oaks, CA: Sage.

Hemker, B. T., Sijtsma, K., Molenaar, I. W., & Junker, B. W. (1997). Stochastic ordering using the latent trait and the sum score in polytomous IRT models. *Psychometrika, 62*, 331-347.

Hemker, B. T., Van der Ark, L. A., & Sijtsma, K. (2001). On measurement properties of continuation ratio models. *Psychometrika, 66*, 487-506.

Hoel, P. G. (1984). *Introduction to mathematical statistics.* Singapore: Wiley.

Hoijtink, H. (1998). Constrained latent class analysis using the Gibbs sampler and postrior predictive p-values: Applications to educational testing. *Statistica Sinica, 8*, 691-711.

Hoijtink, H., & Molenaar, I. W. (1997). A multidimensional item response model: Constrained latent class analysis using Gibbs sampler and posterior predictive checks. *Psychometrika, 62*, 171-189.

Holland, P. W., & Rosenbaum, P. R. (1986). Conditional association and uni-

dimensionality in monotone latent variable models. *The Annals of Statistics, 14*, 1523–1543.

Hollander, M., Proschan, F., & Sethuraman, J. (1977). Functions decreasing in transposition and their applications in ranking problems. *The Annals of Statistics, 5*, 722-733.

Hughes, L. F., Perkins, K., Wright, B. D., & Westrick, H. (2003). Using a Rasch scale to characterize the clinical features of patients with a clinical diagnosis of uncertain, probable, or possible Alzheimer disease at intake. *Journal of Alzheimer's Disease, 5*, 367-373.

Huynh, H. (1994). A new proof for monotone likelihood ratio for the sum of independent Bernoulli random variables. *Psychometrika, 59*, 77-79.

Ip, E. H. (2001). Testing for local dependency in dichotomous and polytomous item response models. *Psychometrika, 66*, 109-132.

Jansen, B. R. J., & Van der Maas, H. L. J. (1997). Statistical test of the rule assessment methodology by latent class analysis. *Developmental Review, 17*, 321-357.

Jasra, A., Holmes, C. C., Stephens, D. A. (2005). Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statistical Science, 20*, 50-67.

Junker, B. W. (1993). Conditional association, essential independence and monotone unidimensional item response models. *The Annals of Statistics, 21*, 1359-1378.

Junker, B. W. & Sijtsma, K. (2000). Latent and monifest monotonicity in item response models. *Applied Psychological Measurement, 24*, 65-81.

Junker, B. W., & Sijtsma, K. (2001). Nonparametric item response theory in action: An overview of the special issue. *Applied Psychological Measurement, 25*, 211-220.

Kang, S. J., Jeong, Y., Lee., B. H., Baek., M. J., Kwon, J. C., Chin, J. & Na, D. L. (2004). How early are initial symptoms recognized in Korean patients with Alzheimer's disease? *International Journal of Geriatric Psychiatry, 19*, 699-700.

Karabatsos, G., & Sheu, C. -F. (2004). Order-constrained Bayes inference for dichotomous models of unidimensional nonparametric IRT. *Applied Psychological Measurement, 28*, 110-125.

Laudy, O., & Hoijtink, H. (2007). Bayesian methods for the analysis of inequality constrained contingency tables. *Statistical Methods in Medical Research, 16*, 123-138.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis.* Boston: Houghton Mifflin.

Ligtvoet, R., Van der Ark, L. A., Bergsma, W. P., & Sijtsma, K. (2009). Polytomous latent scales for the investigation of the ordering of items. Manuscript submitted for publication.

Ligtvoet, R., Van der Ark, L. A., & Sijtsma, K. (2008). Selection of Alzheimer symptom items with manifest monotonicity and manifest invariant item ordering. In K. Shigemasu, A. Okada, T. Imaizumi, & T. Hoshino (Eds.), *New trends in psychometrics* (pp. 225-234). Tokyo: Universal Academy Press.

Ligtvoet, R., Van der Ark, L. A., Te Marvelde, J. M., & Sijtsma, K. (in press). Investigating an invariant item ordering for polytomously scored items. *Educational and Psychological Measurement.*

Loevinger, J. (1948). The technic of homogeneous tests compared with some aspects of "scale analysis" and factor analysis. *Psychological Bulletin, 45*, 507-529.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading MA: Addison-Wesley.

Masters, G. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

Masters, G. N., & Wright, B. D. (1997). The partial credit model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory (pp. 101-121).* New York: Springer.

McLachlan, G. J. & Peel, D. (2000). *Finite mixture models.* New York: Wiley.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika, 12,* 153-157.

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement, 25*, 107-135.

Mellenbergh, G. J. (1995). Conceptual notes on models for discrete polytomous item responses. *Applied Psychological Measurement, 19*, 91-100.

Meng, X. L. (1994). Posterior predictive p-values. *The Annals of Statistics, 22*, 1142-1160.

Michalewicz, Z. (1996). *Genetic algorithms + data structures = evolution programs.* Berlin: Springer.

Mokken, R. J. (1971). *A theory an procedure of scale analysis.* The Hague: Mouton/Berlin: De Gruyter.

Mokken, R. J., & Lewis, C. (1982). A nonparametric approach to the analysis of dichotomous item responses. *Applied Psychological Measurement, 6*, 417-430.

Mokken, R. J., Lewis, C., & Sijtsma, K. (1986). Rejoinder to "the Mokken scale: A critical discussion." *Applied Psychological Measurement, 10*, 279-285.

Molenaar, I. W. (1983). *Item steps* (Heymans Bulletin 83-630-EX). Groningen, The Netherlands: University of Groningen, Department of Statistics and Measurement Theory.

Molenaar, I. W. (1991). A weighted loevinger H–coefficient extending Mokken scaling to multicategory items. *Kwantitatieve Methoden, 12 (37)*, 97-117.

Molenaar, I. W. (1997). Nonparametric models for polytomous responses. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 369-380). New York: Springer.

Molenaar, I. W. (2004). About handy, handmade and handsome models. *Statistica Neerlandica, 58*, 1-20.

Molenaar, I. W., & Sijtsma, K. (2000). *User's manual MSP5 for Windows.* Groningen, The Netherlands. iec ProGAMMA.

Muraki, E. (1990). Fitting a polytomous item response model to Likert-type

data. *Applied Psychological Measurement, 14*, 59-171.

Muraki, E. (1992). A generalized partial credit model: applications for an EM algorithm. *Applied Psychological Measurement, 16*, 159-177.

Narayanan A. (1990). Computer generation of Dirichlet random vectors. *Journal of Statistical Computation and Simulation, 36*, 19-30.

Neal, R. (1991). Bayesian mixture modeling. In C. R. Smith, G. J. Erickson, & P. O. Neudorfer (Eds.), *Proceedings of the 11th International Workshop on Maximum Entropy and Bayesian Methods* (pp. 197-211). Dordrecht, The Netherlands: Kluwer.

Raijmaker, M. E. J., Jansen, B. R. J., & Van der Maas, H. L. J. (2004). Rules in peceptual classification. *Developmental Review, 24*, 289-321.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen, Denmark: Nielsen and Lydiche.

Redner, R. A. & Walker, H. F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review, 26*, 195-239.

Richardson, S. & Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society, Series B, 59*, 731-792.

Rosenbaum, P. R. (1984). Testing the conditional independence and monotonicity assumptions of item response theory. *Psychometrika, 49*, 425-435.

Rosenbaum, P. R. (1987a). Probability inequalities for latent scales. *British Journal of Matematical and Statistical Psychology, 40*, 157-168.

Rosenbaum, P. R. (1987b). Comparing item characteristic curves. *Psychometrika, 52*, 217-233.

Samejima, F. (1969). Estimation of latent trait ability using a response pattern of graded scores. *Psychometrika Monograph,* No. 17.

Samejima, F. (1997). Graded response model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85-100). New York: Springer.

Scheiblechner, H. (1995). Isotonic ordinal probabilistic models (ISOP). *Psychometrika, 60*, 281-304.

Scheiblechner, H. (2003). Nonparametric IRT: Testing the bi-isotonicity of Isotonic Probabilistic Models (ISOP). *Psychometrika, 68*, 79-96.

Shaked, M., & Shantikumar, J. G. (1994). *Stochastic orders and their applications.* San Diego, CA: Academic Press.

Shealy, R. T., & Stout, W. F. (1993). An item response theory model for test bias and differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 197-240). Hillsdale, NJ: Erlbaum.

Sijtsma, K., & Hemker, B. T. (1998). Nonparametric polytomous IRT models for invariant item ordering, with results for parametric models. *Psychometrika, 63*, 183-200.

Sijtsma, K., & Junker, B. W. (1996). A survey of theory and methods of invariant item ordering. *British Journal of Mathematical and Statistical Psychology, 49*, 79-105.

Sijtsma, K., & Meijer, R. R. (1992). A method for investigating the intersection of item response functions in Mokken's nonparametric IRT model. *Applied Psychological Measurement, 16*, 149-157.

Sijtsma, K., & Meijer, R. R. (2007). Nonparametric item response theory and special topics. In C. R. Rao & S. Sinharay (Eds.), *Handbook of Statistics, 26*, (pp. 719-746). Amsterdam: Elsevier.

Sijtsma, K. & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory.* Thousand Oaks, CA: Sage.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B, 64*, 583-639.

Stephens, M. (1997). Bayesian methods for mixtures of normal distributions. *PhD Thesis.* University of Oxford.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society, Series B, 62*, 795-809.

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association, 82*, 528-540.

Tutz, G. (1990). Sequential item response models with an ordered response. *British Journal of Mathematical and Statistical Psychology, 43*, 39-55.

Ünlü, A. (2007). Nonparametric item response theory axioms and properties under nonlinearity and their exemplification with knowledge space theory. *Journal of Mathematical Psychology, 51*, 383-400.

Ünlü, A. (2008). A note on monotone likelihood ratio of the total score variable in unidimensional item response theory. *British Journal of Mathematical and Statistical Psychology, 61*, 179-187.

Van der Ark, L. A. (2001). Relationships and properties of polytomous item response theory models. *Applied Psychological Measurement, 25*, 273-282.

Van der Ark, L. A. (2007). Mokken scale analysis in R. *Journal of Statistical Software, 20*(11), 1-19.

Van der Ark, L. A., & Bergsma, W. P. (2006). Investigating item orderings in polytomous test data. *Unpublished Manuscript.*

Van der Ark, L. A., Croon, M. A., Sijtsma, K. (2008). Mokken scale analysis for dichotomous items using marginal models. *Psychometrika, 73*, 183-208.

Van der Ark, L. A., Hemker, B. T., Sijtsma, K. (2002). Hierarchically related nonparametric IRT models, and practical data analysis methods. In G. A. Marcoulides & I. Moustaki (Eds.), *Latent variable and latent structure models* (pp. 41-62). Mahwah, NJ: Erlbaum.

Van der Ark, L. A., & Van Diem, M. H. P. (2003). Investigating cross-cultural differences in crying using manifest invariant item ordering. In Yanai, H., Okada, A., Shigemasu, K., Kano, Y., & Meulman, J. J. (Eds.) *New developments on psychometrics* (pp. 305-312). Tokyo: Springer.

Van Engelenburg, G. (1997). *On psychometric models for polytomous items with ordered categories within the framework of item response theory.* Unpublished doctoral disseratation. University of Amsterdam.

Van Onna, H. J. M. (2002). Bayesian estimation and model selection in ordered latent class models for polytomous items. *Psychometrika, 67*, 519-538.

Van Schuur, W. H. (2003). Mokken scale analysis: between the Guttman scale and parametric item response theory. *Political Analysis, 11*, 139-163.

Vermunt, J. K. (1999). A general non-parametric approach to the analysis of ordinal categorical data. *Sociological Methodology, 29*, 197-221.

Vermunt, J. K. (2001). The use of restricted latent class models for defining and testing nonparametric and parametric item response theory models. *Applied Psychological Measurement, 25*, 283-294.

Verweij, A. C., Sijtsma, K., & Koops, W. (1999). An ordinal scale for transitive reasoning by means of a deductive strategy. *International Journal of Behavioral Development, 23*, 241-264.

Vingerhoets, A. J. J. M., & Cornelius, R. R. (2001). *Adult Crying. A Biopsychosocial Approach.* Brunner-Routledge, Hove England.

Vorst, H. C. M. (1992). [Responses to the Adjective Checklist] Unpublished raw data.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications.* Cambridge University Press.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24*, 185-201.

Watson, R., Deary, I., & Shipley, B. (2008). A hierarchy of distress: Mokken scaling of the GHQ-30. *Psychological Medicine, 38*, 575-579

Wechsler, D. (2003). *Wechsler Intelligence Scale for Children* (4th ed.). San Antonio, TX: The Psychological Corporation.

Zeger, S. L., & Karim, M. R. (1991). Generalized linear models with random effects; a Gibbs sampling approach. *Journal of the American Statistical Association, 86*, 79-86.

Zhang, J., & Stout, W. F. (1999a). Conditional covariance structure of generalized compensatory multidimensional items. *Psychometrika, 64*, 129-152.

Zhang, J., & Stout, W. F. (1999). The theoretical DETECT index of dimensionality and its application to approximate simple structures. *Psychometrika, 64*, 213-249.

Summary

An invariant item ordering (IIO) means that the items of a test have the same ordering, according to their difficulty or attractiveness, for all subjects for who the test was intended. It also means for each subject, that the scores on the items are expected to be lower as the items become more difficult. IIO facilitates the interpretation of individual test performance and the comparability of test performances of different subjects. However, IIO is also a strong assumption and difficult to satisfy in many tests. As only few item response theory models for analyzing test data imply IIO, only a few methods and models are available to assess whether or not IIO holds. In this thesis, several new models are developed that imply IIO, and methods for assessing IIO holds are generalized to facilitate the research for a wide range of practical applications.

In Chapter 2, method *manifest IIO* was proposed for polytomously scored items. This method assesses whether or not pairs of item response functions for polytomously scored items intersect. Based on the outcome of method manifest IIO, inferences about IIO can be made for both the entire test or for sets of items. If it is concluded that IIO holds for a (subset of items from the) test, then the generalized polytomous coefficient H^T expresses the accuracy of this item ordering. A data example illustrates the applicability of method manifest IIO in combination with coefficient H^T. Tentative rules of thumb are suggested to interpret the accuracy of the item ordering as expressed by coefficient H^T.

In Chapter 3, different items are allowed to have different numbers of score categories. For such items, the combination of monotone increasing item response functions and IIO is investigated simultaneously. In this chapter, par-

ticular emphasis is given to high stakes testing, where the interpretation of the ordering of items has a serious impact on the decisions made about subjects. Two methods are investigated that allow for monotone increasing nonintersecting item response functions to be selected from a test, and the procedure for IIO research is illustrated by an application of the methods to data collected by means of an Alzheimers' symptoms checklist.

In Chapter 4, an approach for testing monotonicity and IIO by means of constrained latent class models is explored, where the latent variable is approximated by a small number of discrete latent classes. The Gibbs sampling procedure is used to estimate the parameters of the ordinal constrained latent class models. The constraints correspond to the assumptions of monotonicity and IIO. Posterior checks are used to identify the items that do not agree with the constraints corresponding to monotonicity and IIO. The procedure for selecting items from the test that satisfy the imposed restrictions is illustrated using a real data set.

In Chapter 5, the perspective on IIO is such that item response functions are required to be distinguishable in the data before the conclusion of IIO is drawn. For realistic sample sizes, it is shown that no more than six items can be assumed to fulfill strict IIO. Another result is that for IIO research most subjects need to be sampled from the extremes of the latent variable distribution, which is where the fewest observations are located. The latter finding explains the need for large samples to be able to reliably establish IIO.

In Chapter 6, a family of *latent scales* for polytomously scored items is defined. Latent scales are item response theory models that imply IIO. The different latent scales are shown to be hierarchically related, and for different levels of the hierarchy testable consequences are derived, allowing the assessment of different definitions of item difficulty ordering. The methodology of Chapter 2 is used to select subsets of items that satisfy a particular latent scale from larger sets. Two data examples illustrate the viability of the approach.

Samenvatting

Een invariante item ordening (IIO) betekent dat de items van een test dezelfde ordening hebben met betrekking tot de moeilijkheden van de items, voor alle personen voor wie de test bedoeld is. Het betekent ook dat voor elk persoon de scores op de items naar verwachting afnemen als de items moeilijker worden. Als zodanig maakt IIO de interpretatie van individuele testprestaties en de vergelijkbaarheid van testprestaties van verschillende personen makkelijker. Echter, IIO is ook een strenge aanname waar moeilijk aan is te voldoen. Aangezien slechts weinig modellen voor het analyseren van testgegevens IIO impliceren, is er nog weinig bekend over IIO en zijn er slechts enkele methoden en modellen beschikbaar om IIO te onderzoeken. In dit proefschrift worden verschillende nieuwe modellen ontwikkeld die wel IIO impliceren en worden methoden voor IIO onderzoek gegeneraliseerd naar een breder scala aan praktische toepassingen.

In hoofdstuk 2 wordt methode *manifest IIO* voorgesteld voor het analyseren van polytoom gescoorde items. Deze methode inspecteert voor paren van item response functies of ze snijden. Op basis van de resultaten van methode *manifest IIO*, kunnen gevolgtrekkingen over IIO gemaakt worden voor zowel de gehele test als voor een verzameling van items. Als wordt geconcludeerd dat IIO geldt voor een (deelverzameling van items uit de) test, dan geeft de gegeneraliseerde polytome coëfficiënt H^T een uitdrukking voor de nauwkeurigheid van deze itemordening. Een datavoorbeeld illustreert de toepasbaarheid van de methode manifest IIO in combinatie met de coëfficiënt H^T. Voorlopig vuistregels worden voorgesteld voor de interpretatie van de nauwkeurigheid

van de itemordening, zoals uitgedrukt in coëfficiënt H^T.

In hoofdstuk 3 worden items beschouwd met verschillende aantallen score-categorieën. Voor dergelijke items wordt tegelijkertijd gekeken naar monotone stijging van de item response functies en IIO. In dit hoofdstuk wordt met name aandacht besteed aan tests waarbij de interpretatie van de volgorde van de items grote invloed heeft op de beslissingen over personen. Twee methoden worden onderzocht die het mogelijk maken om monotone toenemende niet-snijdende item response functies uit een test te selecteren en de procedure voor IIO onderzoek wordt geïllustreerd door een toepassing op symptomen van de ziekte van Alzheimer.

In hoofdstuk 4 wordt een aanpak voor het toetsen van monotonie en IIO onderzocht door middel van gerestricteerde latente klassenmodellen, waarbij de latente variabele wordt benaderd door een klein aantal latente klassen. De Gibbs sampler wordt gebruikt om ordinale restricties op te leggen aan de latente klassenmodellen, waarbij de restricties overeen komen met de veronderstellingen van monotonie en IIO. A posteriori controles worden gebruikt om de items die niet overeenkomen met de aannamen van monotonie en IIO te identificeren.

In hoofdstuk 5 wordt IIO bekeken vanuit het perspectief dat item response functies onderscheidbaar moeten zijn alvorens uitsluitsel te geven over IIO. Voor een realistische steekproefomvang wordt aangetoond dat men een strikte vorm van IIO niet zomaar kan aannemen voor meer dan zes items. Een ander gevolg is dat voor IIO-onderzoek de meeste personen nodig zijn uit de staarten van de latente verdeling waar de minste personen zich bevinden. Deze laatste bevinding verklaart de noodzaak van grote steekproeven om IIO op betrouwbare wijze vast te kunnen stellen.

In hoofdstuk 6 wordt een familie van *latente schalen* gedefineerd voor polytoom gescoorde items. Latent schalen zijn item response theorie modellen die IIO impliceren. De verschillende latente schalen blijken hiërarchisch gerelateerd te zijn en voor verschillende niveaus van de hiërarchie worden toetsbare consequenties afgeleid, waardoor de latente schalen empirisch onderzocht kun-

nen worden. De methode van IIO-onderzoek uit hoofdstuk 2 wordt gebruikt voor het selecteren van items die aan een bepaalde latente schaal voldoen. Twee voorbeelden illustreren de procedure voor het onderzoek naar de latente schalen.